Graph-theoretic multivariate nonparametric procedures

Cortina Borja, Mario Jose Francisco

Award date:
1992

Awarding institution:
University of Bath

Graph-theoretic Multivariate Nonparametric Procedures

Submitted by
Mario José Francisco Cortina Borja
for the degree of PhD
of the
University of Bath

1992

M.J.F. Cortina Borja

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

TO MARION AND MY PARENTS

And if the world were black or white entirely
    And all the charts were plain
Instead of a mad weir of tigerish waters,
    A prism of delight and pain,
We might be surer where we wished to go
    Or again we might be merely
Bored but in brute reality there is no
    Road that is right entirely.

Louis MacNeice

Graph-theoretic Multivariate Nonparametric Procedures

Summary

The difficulties of extending the concept of ordering for multivariate data are responsible for the lack of multivariate generalizations of some well known nonparametric tests. An approximation to an ordered list for the multivariate case consists of linking each data point to other individuals which are regarded as being near to it. There are, of course, many possibilities for deciding when any pair of observations are considered to be near enough so that they may be linked. Our interest is to investigate the adequacy of procedures based on graphs in hypothesis testing and initial analysis of multivariate observations.

Chapter 1 presents the elements of graph theory which appear in the following chapters. In Chapter 2 we study a generalization of the multivariate runs test to the K-sample case and discuss some approximations to its null distribution. In Chapter 3 we explore multivariate tests of hypotheses based on ranks. We use two multivariate ranking methods proposed by Friedman and Rafsky (1979) together with several univariate rank tests. We also describe the approach advocated by Puri and Sen (1971) for constructing multivariate rank tests. Chapter 4 describes tests based on contingency tables based on partitioning the graph nodes. We also discuss the usefulness of such tables in initial multivariate data analysis. Chapter 5 has some examples of the power of the tests described in the previous three chapters against location and scale alternative hypotheses. Chapter 6 examines two measures of multivariate association and prediction and discusses to what extent the normal approximations previously proposed hold for small sample sizes. In Chapter 7 we demonstrate the usefulness of the procedures studied in practice by analyzing three real data sets. In the last chapter we present some concluding remarks and outline several lines for further research.

Acknowledgements

I would like to thank my supervisor, Dr Tony Robinson, for his guidance and encouragement through this work, and all the members of staff and postgraduates from the School of Mathematical Sciences, University of Bath. Also to the Instituto de Investigaciones Antropológicas, Universidad Nacional Autónoma de México, for providing the motivation and the opportunity for doing this course.

The research was funded by the Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México, through a scholarship and the Committee of Vice-Chancellors and Principals of the United Kingdom, through an Overseas Research Studentship Award, and I am grateful for their financial support.

I would like to express my gratitude to the late Francisco Aranda, Glenn Stone, Guy Nason, Howard Grubb, Jaime Litvak, John Eales, Julian Stander, Katerina Tzioli, Merrilee Hurn, Sean Finnigan, Sergio Pezzulli, Stephen Gourley, Wei Liu, William Fulton, and many other colleagues and friends for their help and encouragement, to my family for their unfailing support, and to Marion for everything.

Contents

Chapter 1

Introduction

1.1 Preliminaries

The problems considered in this thesis arise from the question *how can we generalize K-sample nonparametric tests for multivariate data?* Such generalizations may be an important tool whenever the usual assumptions (e.g. normality and homogeneity of variance-covariance matrices) required for parametric multivariate tests are not satisfied by the data. Besides, often due to small sample sizes, in many cases it is not possible to establish accurately to what extent the data fit the said assumptions or if it is adequate to use a large sample approximation to the null distribution used for the parametric test statistic. In these situations, which are frequently found in the practice, the usefulness of a nonparametric alternative is evident.

The literature on nonparametric multivariate tests is scarce. Books concerning either nonparametric statistics or multivariate analysis usually do not mention the subject at all —with some exceptions, for instance, the texts by Du Toit et al. (1986), and Kzranowski (1988), which devote a few pages to it.

The book by Puri and Sen (1971), *Nonparametric methods in multivariate analysis* is practically the only text on the subject. Its central idea is to construct nonparametric tests whose only difference with the usual likelihood ratio tests based on normal theory lies in substituting the observation matrix by a rank matrix —the ranks being determined within each variable. This approach assumes the absence of ties in the

observations. Thus the tests derived from it are not an adequate tool for non-continuous data, in whose analysis an important role is played by univariate nonparametric techniques.

Having this kind of problems in mind it is natural to look for tests which would use dissimilarity measures, as these can be defined in very general terms for a wide variety of data. Mielke et al. (1976, 1981), Mielke (1978, 1979), Berry and Mielke (1983, 1984), and Berry, Mielke and Wong (1986) have extensively studied an approach, known as *Multi-Response Permutational Procedures, MRPP*, which advocates the use of statistics based on weighted sums of distances. The reason for using the term *Permutational* in the name is that the *MRPP* are distribution free, and so their exact permutational null distribution may always be calculated. Of course, this quickly becomes impossible as the sample sizes grow, even for quite small sample sizes. The convergence to a distribution which is easy to calculate is usually very slow, so Mielke and coworkers, in a long series of articles, have proposed several approximations to the permutational distributions of their statistics by matching their first three or four moments. Besides, the *MRPP* have the drawback of requiring the use weights for the distances involved in the test statistics. It is not clear which sort of weights would be more adequate for any particular situation.

Another approach, which contains the *MRPP* as particular cases, is the *Generalized Correlation Coefficients (GCC)* theory. The seminal work in this area is the paper by Daniels (1944). A thorough account of this theory appears in Kendall's book *Rank Correlation Methods* (1962).

Knox (1963) used graph theory to define a generalized correlation coefficient in order to study the space-time clustering of children with leukemia in the North of England. Barton and David (1966) outlined a general method for computing the moments of statistics based on the number of edges which are common to two graphs based on different distance matrices. Much before, Moran (1948) had addressed essentially the same problems.

Figure 1-1 shows the basic idea of Knox's work. The temporal graph has its links defined by pairs of cases detected within a certain time interval of each other; the spatial

configuration shown could correspond to the spread of an epidemic within a region.

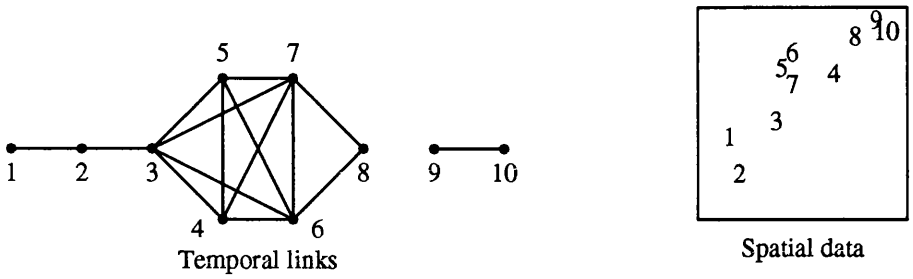


Figure 1-1: An example of space-time clustering

Mantel (1967) and Mantel and Valand (1970) proposed a general permutation statistic, enhanced Barton and David's methods, and obtained explicit expressions for the third and fourth moments of a particular space-time clustering statistic.

We focus our attention on constructing multivariate nonparametric procedures using graph theoretic concepts within the framework of *GCCs* theory. This is an adequate field to work in order to construct all-purpose multivariate tests.

We would like our procedures to have the following features:

1. they should work for virtually any kind of multivariate data
2. they should have distributions that can be easily approximated, in particular for small sample sizes
3. they should give a further insight of the relations amongst the multivariate observations
4. they should be easy to calculate, with a minimum of numerical problems.
5. they should have good power against a wide range of alternative hypotheses.

Our first point of reference are the papers written by JH Friedman and LC Rafsky (1979, 1981 and 1983). These authors applied graph theory concepts to some multivariate analysis problems. Their areas of interest may be summed up as follows:

1. **Hypothesis testing.**

   Friedman and Rafsky gave generalizations of some well known nonparametric tests (Wald-Wolfowitz runs test, Smirnov test) concerning the null hypothesis of homogeneity of two populations.

   Many univariate nonparametric tests are based on the ranks of the pooled sample. The ordered list of observations can be seen as one in which each of its elements is contiguous to points which should have similar values to its own. For multivariate data the construction of a similar relation amongst the observations can be achieved by calculating the distance between every pair of points in order to link those pairs which are "near". Statistics conditioned on the observed links would lead to nonparametric tests. Whatever the criterion is for selecting a set of neighbours, the notion of linking pairs of points corresponds to a graph; thus, graphs are a natural tool for generalizing well known nonparametric tests to the multivariate case.

2. **Planing.**

   Sometimes a planar or tridimensional representation of multivariate observations is desirable. As a follow-up technique to their two-sample tests, Friedman and Rafsky (1981) introduced a two-dimensional mapping technique called *planing*. Many techniques used to obtain low dimensional representations make use of projections algorithms. Such is the case of plots based on the few first principal components or on projection pursuit methods. They may produce very good results evaluated either in terms of an overall discrepancy measure (stress) between the original distance matrix and the distance matrix obtained for the low dimensional configuration, or by means of the proportion of total variance contained in the data projections on the first few principal components. However, these low dimensional configurations can be computationally expensive to obtain and may be affected by numerical difficulties.

   Another possibility would be to use multidimensional scaling techniques e.g. those proposed by Kruskal (1977) or non linear mappings like the one originally

proposed by Sammon (1969). In this case, one attempts to construct a planar representation of the observations that minimizes a stress function. Instead of looking for configurations such as those indicated above, the *planing* technique of Friedman and Rafsky (1981) constructs a two-dimensional configuration that preserves exactly, by means of a triangulation method (Lee *et al.*, (1977)), only a few of the distances. For p-dimensional configurations it is possible to preserve the distances amongst $p + 1$ points. Such distances are chosen from those that define links on a certain class of graphs, and the order in which the points are plotted can be used to highlight the distance relations existing in the data with respect to some particular point.

3. **Probability-Probability Plots.**

   Another use of the list of ranked univariate observations for the pooled sample is the construction of P-P plots. In the univariate two-sample case, if X and Y denote the values for each sample, a P-P plot for samples of sizes n_1 and n_2 is a plot of the $n_1 + n_2$ points $(\hat{F}_X(z), \hat{F}_Y(z))$, where $\hat{F}_X$ and $\hat{F}_Y$ are the empirical distribution functions and z takes values over the pooled sample; the points are connected in order of increasing z. If the samples were identical, then the plot would be like the graph of $f(x) = x$; location differences would be reflected by plots lying predominantly above or below that line and scale differences would tend to produce plots that lie on either side of that line, cross it and remain on the other side for the remainder of the plot. Again, a possibility for generalizing the concept of an ordered sequence of values to the multivariate case, is to link points which are "near" in the multidimensional space and has been explored by Friedman and Rafsky (1981).

4. **Multivariate measures of association and prediction.**

   Daniels (1944) defined a generalized correlation coefficient between two data matrices **X** and **Y** as

$$\Gamma = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}\, b_{ij} \tag{1.1}$$

where a_{ij} and b_{ij} are scores determined by the values of the data points on **X** and **Y**, respectively and N is the sample size. Examples of these statistics that may be expressed as generalized correlation coefficients include Pearson's r, Spearman's ρ and Kendall's τ, which correspond to the scores choices $a_{ij} = x_i - x_j$, $b_{ij} = y_i - y_j$; $a_{ij} = \text{rank}(x_i) - \text{rank}(x_j)$, $b_{ij} = \text{rank}(y_i) - \text{rank}(y_j)$ and $a_{ij} = \text{sign}(x_i - x_j)$, $b_{ij} = \text{sign}(y_i - y_j)$, respectively.

It is possible to use Γ for multivariate observations, as the scores a_{ij}, b_{ij} can be defined for vectors. But, once more, one needs some multivariate analogy to the concept of ranking which defines correlation coefficients like ρ and τ. Friedman and Rafsky (1983) studied two multivariate correlation coefficients. They are suitable for assesing the significance of the correspondences between two data matrices.

In this thesis we are concerned principally with the first and last points of those covered by Friedman and Rafsky, as our central interest is to address the problem of hypothesis testing for multivariate observations in a nonparametric context.

Graph theory is a convenient framework to present and study relations amongst multivariate observations. Most of its applications in statistics are based on graphs whose nodes represent the sample points and whose edges link individuals which are "near" to each other. The relation of "closeness" is defined with respect to a dissimilarity (or similarity) matrix calculated on the multivariate observations. Every edge of the graph has an associated weight which is usually a monotone function of the dissimilarity between the pair of nodes defining the edge.

Most of the applications of graph theory in statistics have been in cluster analysis. It is possible to provide rigorous definitions of a cluster using graphs. For instance, consider the construction of dendrograms. First one obtains a graph using the nodes to represent the individuals under study, and defining an edge in it if a pair of nodes

satisfies some relation of interest based on a distance matrix, e.g., if one of them is the nearest neighbour of the other. Then, it is possible to proceed to group subsets of the data in such a way that the distances between pairs of individuals in a group are always less or equal than some threshold distance, which corresponds to a certain edge in the graph.

Figure 1-2 illustrates these ideas. For the dendrogram shown there, the threshold distances are defined by the following pairs of nodes: (2,3), (2,4), (3,5), (3,1). This sequence of pairs of nodes corresponds to the *single linkage method*, in which the threshold distance between two groups corresponds to the shortest distance amongst those defined by pairs of nodes from different groups. If instead we define the threshold distances to be those which are the largest distances between pairs of nodes from different groups, then the sequence defining the dendrogram would be (2,3), (3,4), (2,5) and (4,1). These procedures consider N groups with 1 individual each and end up with 1 group with N individuals within it.

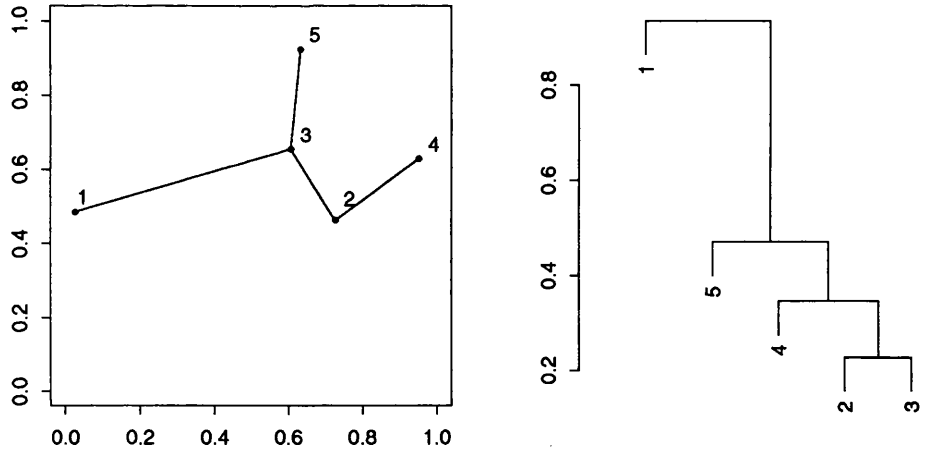


Figure 1-2: A dendrogram based on a graph

The book by Jardine and Sibson (1971), and the papers by Ling (1972), Matula (1972) and Hubert (1974), amongst others, provide a theoretical basis for cluster analysis.

The contribution of Friedman and Rafsky (1979) is to stress that graphs with edges

defined by pairs of points which are "near" can be used to generalize ideas like the ranking of univariate data to the multivariate case. The use of the ranks obtained from these graphs have received little attention. The number of papers with specific applications of the Friedman-Rafsky multivariate runs tests is very small. They have been applied in the analysis of a clinical diagnostic classification process by Brohet et al. (1984). Smith and Jain (1984) used these tests to provide a test for uniformity for multivariate data. Karlin et al. (1983), Seber (1984) and Miller (1985) mention the Friedman-Rafksy tests as an alternative approach useful to analyze multivariate data with complex structure.

There is, however, some evidence that the tests thus derived may have good power against a wide range of alternatives in various dimensions. Friedman and Rafsky (1979) studied the performance of tests based on sequences of orthogonal spanning trees and stated their usefulness for multivariate normal and nonnormal data. Whaley (1983) showed that these tests are equivalent to the procedures derived by Cliff and Ord (1973) for spatial autocorrelation tests and to the *MRPP* proposed by Mielke et al. (1976).

Whaley and Quade (1985) used graphs which have an edge defined if and only if the distance between any pair of points is less than some specific threshold value. They found that, in some cases, the power of such nonparametric tests can be as high as the one obtained with parametric tests. A drawback of this approach is that it is not clear how to choose an optimal threshold value to define the graphs used by Whaley and Quade. However, they showed that the runs tests thus defined may have greater power than Hotelling's T^2 test. They also compared the threshold-graph tests with Friedman-Rafsky's original tests and found that, in general, the latter have better power.

Schilling (1986) considered two-sample tests based on the nearest neighbours from a Euclidean distance matrix. He worked with the weighted proportion of all n-nearest neighbours in which observations and their neighbours belong to the same sample. He found that for several alternatives his tests' performances were similar to that of the Friedman and Rafsky's tests. Henze (1988) extended Schilling's results to any distance generated by a norm on $\mathbb{R}^p$. Although both authors obtained expressions for

their optimal weights, it is not clear how these would be generalized to the K-sample case.

Friedman and Rafsky, Whaley and Quade, Schilling and Henze worked only with graphs derived from nearest neighbours graphs and from minimum spanning trees and left aside a few extensions of their ideas. We explore the application of a wider range of graphs to enhance their approach and study in greater detail some points mentioned by those authors. Of particular interest is the construction of small sample approximations to the null distribution of the statistics discussed in the Friedman-Rafksy papers. The asymptotic normality for statistics of the form (1.1) was proved by Friedman and Rafsky (1983) for very general conditions using the results of Daniels (1944) on the distribution of generalized correlation coefficients over the space of sample permutations.

In the rest of this chapter we introduce some graph theory concepts and discuss some graphs and algorithms that will be used throughout the following chapters.

1.2 Graph Theory Concepts

The terminology and notation in graph theory texts are far from being uniform, so we devote this section to present the graph-theoretic machinery used in the rest of the thesis.

A *graph* $\mathcal{G}$ is an ordered pair (V, E), consisting of a finite, non empty set $V(\mathcal{G})$ of *vertices* (also called *nodes*) and a (possibly empty) set $E(\mathcal{G})$ of *edges*, whose elements are defined by pairs of vertices. We say that an edge *links* the two nodes defining it and that it is *incident* on both of them; a node is *adjacent* to all its incident edges. Figure 1-3 shows an example of a graph.

In this section, $\mathcal{G}$ and $\mathcal{H}$ denote graphs; $V(\mathcal{G})$ and $E(\mathcal{G})$ are the node set and the edge set, respectively. A *path* between any two different nodes $v_1, v_2 \in V(\mathcal{G})$ is an alternating sequence of nodes and edges of $\mathcal{G}$ having v_1 as its first element and v_2 as the last one; the edges in the path must be adjacent and the inner nodes must be different. If $v_1 = v_2$, we call that path a *cycle*. The *length* of a path is the number of edges included

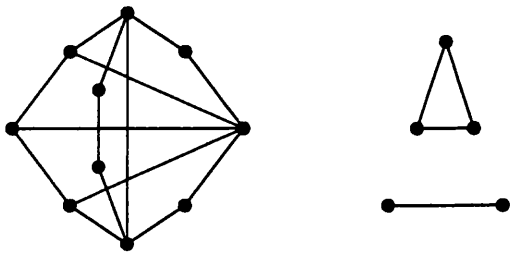


Figure 1-3: A disconnected graph

in it. A *connected graph* has a path between every pair of nodes. The *degree* of a node is the number of edges incident on it.

A graph $\mathcal{H}$ is a *subgraph* of $\mathcal{G}$ if $V(\mathcal{H}) \subseteq V(\mathcal{G})$ and $E(\mathcal{H}) \subseteq E(\mathcal{G})$; if $V(\mathcal{H}) = V(\mathcal{G})$, then $\mathcal{H}$ is a *spanning subgraph* of $\mathcal{G}$. A *complete* graph is one in which every two nodes are adjacent; the complete graph with N nodes, will be denoted by $\mathcal{K}_N$. The *empty Nth* graph, $\mathcal{E}_N$, has N nodes and no edges (Figure 1-4).

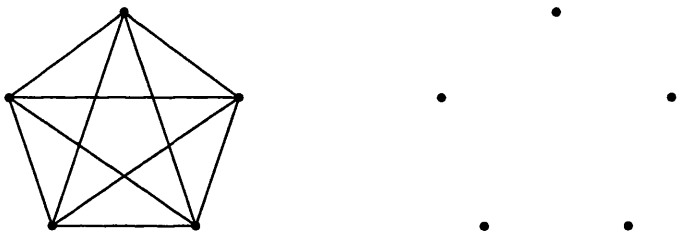


Figure 1-4: $\mathcal{K}_5$ and $\mathcal{E}_5$

A *tree* is a connected graph without cycles. For a graph with N nodes, a tree has $N-1$ edges. An *edge weighted* graph has a real number assigned to each edge. A *minimum spanning tree* (*MST*) of an edge weighted graph is a spanning tree for which the sum of the edge weights is minimum. The *eccentricity* of a node P in a tree is the number of edges in a path with greatest length beginning in that node; the node at the other end of such a path is called an *antipode* of P. The path between a node with largest eccentricity and one of its antipodes is called a *diameter*. A *centre* of the tree is a node

for which the eccentricity is minimum (Figure 1-5). The antipodes of a diameter are called its *ends*.

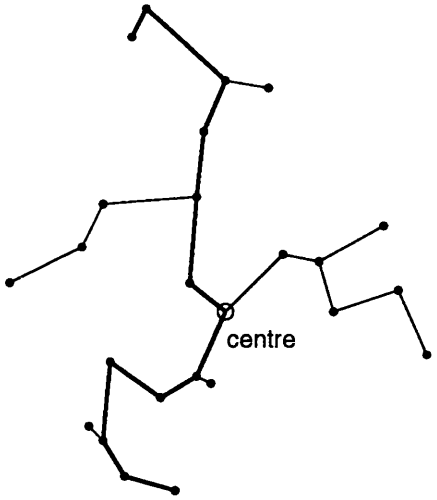


Figure 1-5: A minimum spanning tree, its diameter, and its centre

A *rooted tree* has one of its nodes labelled as its *root* (Figure 1-6). For every node in a rooted tree, its *depth* is the length of the path between it and the root; the *height* of a rooted tree is the maximum depth in it. The *parent* of any node P is the penultimate node on the path beginning with the root and ending with P; the nodes which are different to P in such a path are its *ancestors*. The *daughters* of any node P are those nodes that are not its parents but are linked to it; its *descendants* are all the nodes for which P is an ancestor. To *traverse* a tree is a procedure in which all the nodes are visited according to some order, usually specified in relation with a sequence of rootings.

Two graphs are *orthogonal* if they have the same node set and the intersection of their edge sets is empty. The *complement* of a graph $\mathcal{G}$, denoted by $\bar{\mathcal{G}}$, has node set equal to $V(\mathcal{G})$ and $(u, v) \in E(\bar{\mathcal{G}})$, if and only if $(u, v) \notin E(\mathcal{G})$. The *intersection* and the *union* of two graphs $\mathcal{G}$ and $\mathcal{H}$ which have a common node set are the graphs with the same node set and the edge sets $E(\mathcal{G}) \cap E(\mathcal{H})$ and $E(\mathcal{G}) \cup E(\mathcal{H})$, respectively.

A graph is *planar* if it can be embedded in the plane without crossings, i.e. in a way such that distinct edges intersect only at nodes. A straight line planar embedding of a planar graph determines a partition of the plane called *planar subdivision* or *map*.

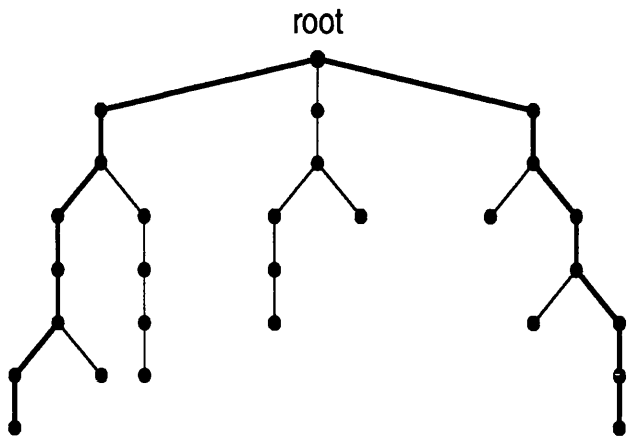


Figure 1-6: A minimum spanning tree rooted at its centre

A representation of a planar graph in the plane is called a *plane* graph. Let v, e, and f denote respectively the number of vertices, edges and regions (including the single unbounded region) of a map. These three parameters are related for a connected plane graph by the classical Euler's formula

$$v - e + f = 2$$

The *incidence matrix* of a graph with N $(N \geq 2)$ nodes is the $N \times N$ matrix $\mathbf{A} = [a_{ij}]$ in which a_{ii} is the degree of node i and $a_{ij} = -1$ if there is an edge defined by the nodes i and j and 0 otherwise. This matrix can be used to study some interesting properties of graphs.

Given a connected graph $\mathcal{G}$ with N nodes and e edges, its *complexity* $c(\mathcal{G})$ is the number of spanning trees contained within $\mathcal{G}$; the matrix tree Theorem (Wilson, 1972) establishes that

$$c(\mathcal{G}) = \alpha(\mathbf{A}) \tag{1.2}$$

where α denotes the cofactor of any of the entries of $\mathbf{A}$. Another useful property of $\mathbf{A}$ is that rank $(\mathbf{A}) = N - s$, where s is the number of connected components within $\mathcal{G}$.

Given N and e, every graph which is possible to form with these parameters can be regarded as a point in a probability space which has $\binom{M}{e}$ elements, where $M = \frac{1}{2}\binom{N}{2}$. Moon (1971, p. 45) showed that the expected value of $c(\mathcal{G})$ over the elements of this space is

$$\mathrm{E}(c(\mathcal{G}_{N,e})) = N^{N-2}\binom{M-(N-1)}{e-(N-1)}\binom{M}{e}^{-1}. \tag{1.3}$$

Moon (1971) also obtained a formula for $\mathrm{var}(c(\mathcal{G}_{N,e}))$, but it is much more difficult to calculate than the mean, and in general, it does not provide useful information in this context, as it is asymptotically of order $\mathcal{O}((\mathrm{E}\,c(\mathcal{G}_{N,e}))^2)$, as Janson (1986) proved. The distribution of the complexity of a graph is not known.

We shall use these concepts in the following chapters. The graphs considered have their nodes corresponding to the data points; the edges are defined by pairs of points, with edge weights determined by a dissimilarity matrix $\mathcal{D} = [d_{ij}]$. Given this set up, the way of defining the edges is determined by the graph in question.

1.3 Nearest Neighbours Graphs

The concept of n-order nearest neighbourhood has been widely used in Statistics: cluster analysis, bivariate splines, image analysis, and spatial statistics are only some fields in which it plays an important role. A definition of *nearest neighbourhood of order n (n-NN)* follows. Let d be any distance function. The *n-NN* to the point x_i is the point x_j such that $d_{ik} < d_{ij}$ for exactly $n-1$ values of k, with $(1 \leq k \leq N)$ and $k \neq i, j$. If we assume that the off-diagonal elements of the distance matrix are all distinct, then the nearest neighbour of every order for each point is unique. In many cases, we can think that ties within the distances occur with probability zero. However if ties do occur, mainly due to rounding or measurement limitations effects, it is possible to handle them in the following manner, suggested by Schilling (1986): suppose that Q observations are equidistant from x_i, with other $n-1$ points strictly closer to x_i; assign a random permutation of the appropriate ranks $n, n+1, \ldots, n+Q-1$ to these Q points

in forming the *NN* list for x_i. The *n*-nearest neighbour graph (*n-NNG*) is obtained by linking the nodes of *i-NN*, with $1 \leq i \leq n$.

The nearest neighbourhood relation is not symmetric. If two points are such that one is the *m*th *NN* of its own *n*th *NN*, then they are called *reflexive nearest neighbours of order* (*m, n*). Pickard (1972), Cox (1981) and Henze (1988) amongst others, provided some interesting results concerning this relation.

Friedman *et al.* (1975) obtained an efficient algorithm to calculate *n-NN* in *p* dimensional spaces; it may be applied to any distance measure. The efficiency of this algorithm depends on the dimensionality of the data as well as on the distance function used. Its authors gave some lower bounds for several distance measures and showed that it compares favourably to the usual brute force type algorithms.

Figure 1-7 shows the first two *NNG* obtained for 50 points in the plane. The number of links of each graph is also shown.

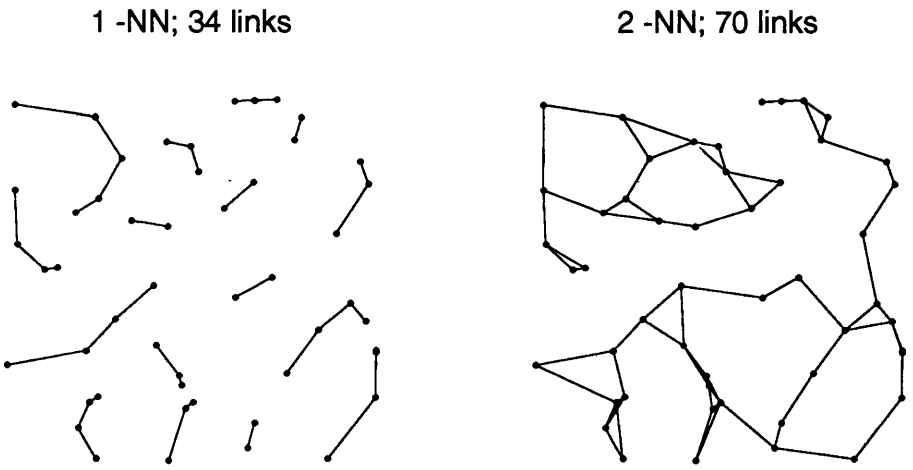


Figure 1-7: Nearest Neighbours Graphs

The n-*NNG* is not necessarily a connected graph; however, for Euclidean distances over points on $\mathbb{R}^p$, as *p* increases, the proportion of connected n-*NNG* also increases for every order of nearest neighbourhood.

1.4 Orthogonal Minimum Spanning Trees

A minimum spanning tree is an edge weighted tree for which the sum of the weights is a minimum over the set of all the spanning trees with a fixed number of nodes. If we take the edge weights to be associated with a distance matrix calculated for the N data points, then an *MST* connects all the nodes with $N-1$ edges and these edges represent pairs of points that are close together.

It is possible to apply the definition of *MST* to construct orthogonal spanning trees. Thus, the 2-*MST* is the union of the 1-*MST* and the spanning tree obtained, if edges from the 1-*MST* are excluded, by minimizing its total length. In general, an *n-MST* is formed as the union of the first $(n-1)$-*MST*s and the *MST* obtained without including any edge belonging to the previous *MST*s.

Two properties of the first *MST* are expressed in the following theorems.

Theorem 1 *1-NNG* $\subseteq$ *1-MST*

Proof: If $N = 2$, the theorem is true; this is the basis of the induction. Now suppose that for $N-1$ nodes, 1-*NNG* $\subseteq$ 1-*MST*; if we add another node, then to be sure that the resulting tree has still minimum sum of edge weights, we must join the new node to its nearest neighbour.

Theorem 2 *If any edge of an MST is deleted, thus dividing the points into two disjoint subsets, then the deleted edge weight corresponds to the smallest interpoint distance between the two subsets.*

Proof: Let $\mathcal{G} = (V, E)$. Let N be the number of nodes in $\mathcal{G}$, and $U \subset V$ be any subset of V; consider e to be an edge of minimum length amongst those edges connecting nodes in U with nodes in $V-U$. To prove the theorem it suffices to show that there is an *MST* which contains e. Let T_0 be an *MST*. If $e \notin T_0$, then add e to T_0, thus forming a cycle that contains e and at least one more edge, e', which connects nodes from U and $V-U$. Deleting e' from $T_0 \cup e$ we obtain another spanning tree, T_1, as it is a connected graph with $N-1$ edges. But length $(e) \leq$ length (e'), and so the total length of T_1 is less or equal to that of T_0, implying that T_1 is an *MST*.

There is a large variety of algorithms to construct *MST*s. The following one was proposed by Kruskal (1956):

1. Sort the elements of the distance matrix $\mathcal{D}$ in ascending order.
2. Follow that sorted list and select edges making sure that no cycle is formed.
3. Stop when $(N-1)$ edges have been selected. They form an *MST*.

Cayley's Theorem (Moon, 1971) establishes that the number of different trees that is possible to construct from N nodes is N^{N-2}. In view of this fact, the construction of the *MST* seems remarkably simple. Prim (1957) obtained a better algorithm, based on two principles for *MST* construction. Let an *isolated node* be a node which has not been linked to the *MST* at some stage of the *MST* construction. A *fragment* is a spanning tree of a subgraph and an *isolated fragment* is a fragment which, at some stage of the construction, has not been linked to the rest of the graph. The *distance* between any node and a fragment of which it is not a member is defined as the miminum interpoint distance between that node and the nodes in the fragment. A *nearest neighbour of a node* is another one whose distance to the specified node is as small as that of any other node. A *nearest neighbour of a fragment* is a node which has a distance as small as that of any other node from the fragment. Prim's principles are:

1. Any isolated node can be linked to a nearest neighbour.
2. Any isolated fragment can be linked to a nearest neighbour by a shortest edge not included in any other fragment.

Using these principles, Prim's algorithm can be enunciated as follows:

1. Begin with any arbitrary node. Link it with one of its nearest neighbours; this constitutes the first fragment tree.
2. Find the smallest interpoint distance between a point that does not belong to any fragment and a point already linked to the *MST*; link this pair of points.
3. Repeat step 2 $N-2$ times.

FORTRAN programs implementing Prim's algorithm for any distance matrix (without ties) appeared in Ross (1969) and Whitney (1972). They both run in $\mathcal{O}(N^2)$ time. Shamos and Hoey (1975) presented an algorithm to obtain *MST*s in the plane which is at most $\mathcal{O}(N \log N)$; it is based on the Delaunay triangulation of the data points. Bentley and Friedman (1975) and Yao (1982), amongst others, proposed algorithms for *MST*s in higher dimensions for which the computation time is, in average, of lesser order than $\mathcal{O}(N^2)$. However, all these algorithms assume that a variety of geometrical properties hold for the data points in order to run in less than $\mathcal{O}(N^2)$ time. In general, for pooled sample sizes of a few hundred points, Prim's algorithm still gives the most efficient and general way of finding *MST*s.

Friedman and Rafsky (1979) gave the following $\mathcal{O}(N)$ algorithm to find a centre of an *MST*:

1. Choose an arbitrary node as root.
2. Find the node in the *MST* of greatest depth; this is an antipode.
3. Choose this antipode as the root and find its antipode.
4. These two nodes form a diameter of the *MST*. With one of them as a root, find a node on the diameter whose depth is as close as possible to half the depth of its antipode; this is a centre of the *MST*.

An algorithm to construct an *n-MST* is as follows:

1. Calculate the 1-*MST*.
2. **do** $n-1$ times
   (a) Assign a value of ∞ to the entries of the distance matrix which correspond to an edge included in previous *MST*s.
   (b) Obtain an *MST* for that modified distance matrix.

Figure 1-8 shows the two first orthogonal *MST*s for 50 points in the plane.

Zahn (1971) proposed a great variety of data analysis techniques using the 1-*MST*; his main interest was to develop methods that define clusters for 2-dimensional data

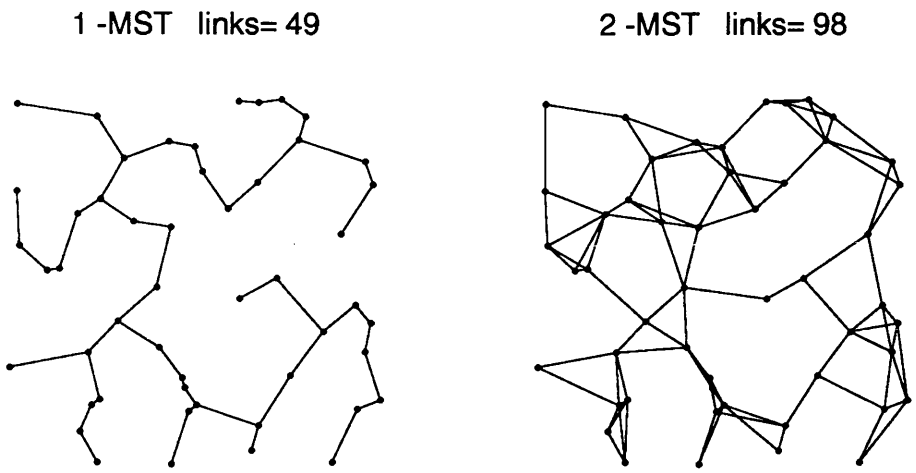


Figure 1-8: Orthogonal Minimum Spanning Trees

sets. Zahn pointed out that edges in the *MST* tend to follow steepest gradients in point density. The centre of the *MST* tends to lie near the geometric centre of the multivariate data points and can be regarded as analogous to the median of a univariate data set. These two remarks mean that, if we root the *MST* at its centre, then for spherically symmetric distributions (and their transformations), the depth of every node is a quantity similar to the distance between every point and the mode of the distribution.

Some problems may occur in the specification of the *n-NNG* or the *n-MST* if there were some pairs of points which have the same distance —usually this possibility is ruled out by assuming that ties occur with probability zero. For the former of these graphs, we can just consider that any point may be allowed to have more than one nearest neighbour of any order.

For the *MST*s, the solution is not as obvious because there might be as many minimal spanning trees as the number of pairs points with the tied distances. However, if this number is relatively small, the possible *MST*s should not differ much with the one finally chosen. Chatfield and Collins (1980, §11.4) mentioned a suggestion made by Sibson in order to deal with this problem. The idea is to define an invariant graph

using the ultrametric distances. This graph coincides with the *MST* if there are no ties, but may have cycles as well as a total weight larger than the one of the *MST*s. This advice is ad hoc for cluster analysis problems. As we will see, in some applications we specifically require the tree structure. Other times, we are mainly interested in orthogonal *MST*s, and these would pick up these distances eventually. In general, unless we were examining the neighbourhood structure as such with an *MST*, we followed the convention of picking the first spanning tree obtained, regardless of the tied distances involved in it. If there are relatively few of them, the results should not be affected very much. Otherwise, we followed Sibson's suggestion.

1.5 Exodic Trees

Gilbert (1964) defined the exodic tree (*ET*) as a "not quite minimal spanning tree". He proposed this graph in order to calculate an upper bound to the total edge length of the minimum spanning tree; we will follow this approach later on. Roberts (1968) enhanced Gilbert's results. The term *exodic* was given to this tree because some paths contained in it radiate outwards from the root. It has not been used in hypothesis testing, where it may increase power against scale alternatives. In another application, it may provide a convenient framework to construct low dimensional representations of multivariate observations in a way that highlights the relationships of the points with respect to any selected location in the sampling window.
An at most $\mathcal{O}(N^2)$ algorithm to obtain *ET*s is as follows:

1. Choose any point as the root of the *ET* and label it as $x_{(1)}$.

2. Label the rest of the points as $x_{(2)}, \ldots, x_{(N)}$ according to the ascending distances between these points and $x_{(1)}$.

3. Link $x_{(i)}$ $(i \geq 2)$ to the point $x_{(j)} \in \{x_{(1)}, x_{(2)}, \ldots, x_{(i-1)}\}$ chosen to minimize the distance between $x_{(i)}$ and $x_{(j)}$.

It is easy to see that this construction leads to a tree: by induction on i, if the edges made at $x_{(2)}, \ldots, x_{(i-1)}$ form a tree, then adding an edge containing $x_{(i)}$ induces no cycles. The

total length of an *ET* is usually not much larger than that of an *MST*, as from all the trees containing paths going from $x_{(1)}$ to $x_{(i)}$, for each $i > 2$ through nodes $x_{(j)}$ with increasing distances with $x_{(1)}$, the *ET* rooted at $x_{(1)}$ has minimal length.

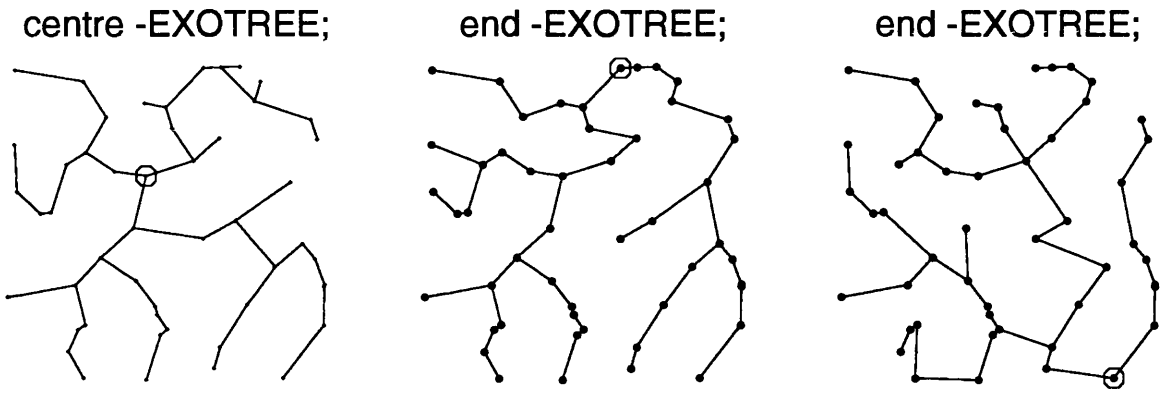


Figure 1-9: Exodic trees rooted at the centre and at the diameter extremes of the *MST*

Figure 1-9 shows three different exodic trees for 50 points in the plane. The first was obtained taking the centre of the corresponding 1-*MST* as the root; the other two proceed from rooting the *ET* in each of the ends of the diameter of the 1-*MST*.

1.6 Relative Neighbourhood Graphs

There are many possibilities for considering two points in the p dimensional space as being "relative neighbours". Using the concept of *relative close neighbours* in the form first proposed by Lankford (1969), Toussaint (1980) investigated some aspects of the relative neighbourhood graph (*RNG*). His definition is as follows:

two points x_i and x_j define an edge of the *RNG* iff

$$d(x_i, x_j) \leq \max_{k \neq i,j} \max [d(x_i, x_k), d(x_j, x_k)],$$

which is equivalent to say that

two points x_i and x_j are not an edge of the *RNG* iff

$$d(x_i, x_j) > \max_{k \neq i,j} \max [d(x_i, x_k), d(x_j, x_k)],$$

Intuitively, this means that two points define an edge in the *RNG* iff they are at least as close to each other as they are to any other point. The following formulation is equivalent:

> Two points x_i, x_j are linked in the *RNG* iff the intersection of the open hyperspheres with radii $d(x_i, x_j)$ centered at x_i and x_j has no point in it.

Toussaint (1980) enunciated the following theorem for a planar configuration. It gives a relation between the 1-*MST* and the *RNG*.

Theorem 3 *1-MST* $\subseteq$ *RNG*

Proof: Let $\mathcal{C}$ denote the interior of the intersection of the spheres with centres at any points x_i, x_j and radii $d(x_i, x_j)$, and let $\mathcal{B}$ be the boundary of $\mathcal{C}$, as shown in Figure 1-10. Suppose any third point x_k lies outside $\mathcal{B}$, i.e., the *MST* is unique; thus, x_k must be either in $\mathcal{C}$ or in $\bar{\mathcal{C}} = (\mathcal{C} \cup \mathcal{B})^c$. If $x_k \in \mathcal{C}$, then $d(x_i, x_k) < d(x_i, x_j)$, and $d(x_j, x_k) < d(x_i, x_j)$, so $\overline{x_i x_j} \notin$ 1-*MST*. This gives a necessary but not sufficient condition for an edge being in the 1-*MST*: all the other points must lie in $\bar{\mathcal{C}}$, and this is a necessary and sufficient condition for $\overline{x_i x_j} \in$ *RNG*, as claimed.

The obvious way of calculating *RNG*s is the $\mathcal{O}(N^3)$ brute force algorithm proposed by Toussaint (1980):

1. Compute the distance matrix $\mathcal{D} = [d_{ij}]$
2. For each pair of points (x_i, x_j)
   (a) compute $d^k_{\max} = \max\{d_{ki}, d_{kj}\}$ for $k = 1, \ldots, n, k \neq i, k \neq j$
   (b) search for a point, x_k, such that $d^k_{\max} < d_{ij}$; if no such x_k is found, then define an edge with points x_i and x_j.

Urquhart (1980) presented an algorithm to obtain *RNG*s which is at least $\mathcal{O}(N^2)$:

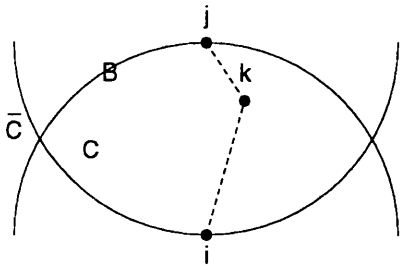


Figure 1-10: 1-*MST* $\subseteq$ *RNG*

1. Compute the distance matrix $\mathcal{D} = [d_{ij}] = [d(x_i, x_j)]$

2. For each pair of points (x_i, x_j) that has not been rejected as an edge of the *RNG*, compute $d^k_{\max} = \max\{d_{ki}, d_{kj}\}$ for $k = 1, \ldots, n, k \neq i, k \neq j$

   (a) if $d_{ij} > d^k_{\max}$ then reject $\overline{x_i x_j}$ as an edge of the *RNG*

   (b) if $d_{ij} < d^k_{\max}$ then reject the pair of points separated by $d^k_{\max}$ as an edge of the *RNG*

   (c) if $d_{ij} \leq d^k_{\max}, \forall k$, then $\overline{x_i x_j} \in RNG$.

Some algorithms which are more efficient for particular cases have been proposed. O'Rourke (1982) discussed algorithms for *RNG*s that run in $\mathcal{O}(N^2 \log N)$ considering the L_1 metric in the plane and the L_∞ metric in higher dimensions. Urquhart (1980, 1982) and Supowit (1983) gave algorithms for Euclidean metrics in the plane that run in $\mathcal{O}(N \log N)$ time. Supowit (1983), proving rigorously the adequacy of his own algorithms, corrected Urquhart's work for this case and presented a review of algorithms for calculating *RNG*s. The chief interest of this author was to investigate in which conditions an *MST* for planar configurations could be obtained in linear time by deleting edges of the *RNG*. In this line, he found an algorithm which runs in $\mathcal{O}(N)$ time for points enclosed within a convex polygon. Supowit (1983) pointed out that the fastest general algorithms for *RNG*s so far known are still those presented by Urquhart (1980, 1982).

Lefkovitch (1985) proposed a generalization to construct higher order *RNG*s which may be expressed as follows:

> the edges on the *n-RNG* ($n > 1$) link points which were not already linked in previous *RNG*s and which have at least one common relative neighbour of lower order.

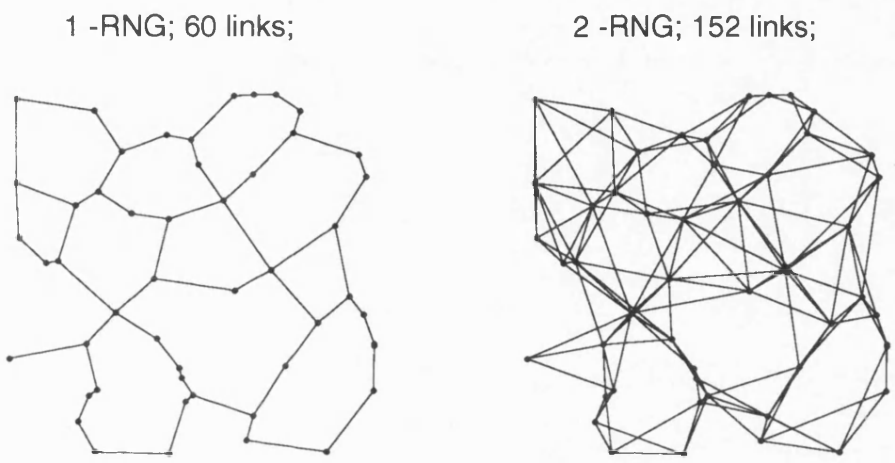


Figure 1-11: Generalized Relative Neighbourhood Graphs

Consider an $N \times N$ boolean matrix $\mathbf{R}$ with value TRUE at entry i-j if the nodes i and j define an edge in the graph and FALSE otherwise. Then, the matrix of this type corresponding to the nth order *RNG* can be obtained by

$$\mathbf{R}_n = \mathbf{R}^2 \wedge \bar{\mathbf{R}} \tag{1.4}$$

where $\mathbf{R}$ is the boolean matrix formed with the union of the first $n-1$ *RNG*s, $\bar{\mathbf{R}}$ denotes its complement and $\wedge$ is the Boolean matrix AND operation. Obviously, $\mathbf{R}_m \wedge \mathbf{R}_n =$ [FALSE], for all $m \neq n$. Figure 1-11 gives an example of this generalization.

Figure 1-12 contains an example of a more qualitative comparison between the *NNG*, the *MST* and the *RNG*. Using Euclidean distances for 100 configurations of uniform

random numbers, it shows the percentage of the total number of nodes which were leaves, i.e. that had degree 1. Clearly, this may change for other distributions.

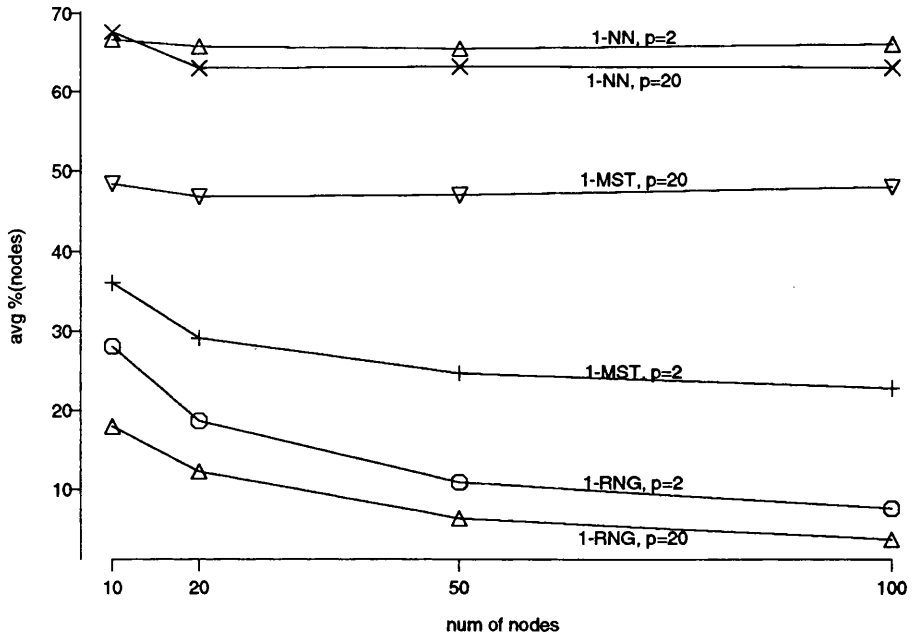


Figure 1-12: % of leaves

1.7 Gabriel Graphs

This graph was first proposed by Gabriel and Sokal (1969) to define connectedness for a set of localities within geographical regions. Preparata and Shamos (1985) mentioned some properties and applications in pattern recognition problems. The definition of a Gabriel graph (*GG*) is as follows:

> two points x_i, x_j are linked in *GG* iff the open hypersphere with diameter $d(x_i, x_j)$ centered at the midpoint of the segment joining x_i and x_j contains no other point.

This is equivalent to say that x_i and x_j define an edge in the *GG* iff

$$d^2(x_i, x_j) \leq d^2(x_i, x_k) + d^2(x_j, x_k), \ \forall k \neq i, j$$

To compute a *GG* from a set of N p-dimensional points, we use Urquhart's (1980, 1982) general algorithm as described in the previous section. The only difference is that, for each pair of points x_i, x_j, instead of testing if there is any other point in the intersection of the two hyperspheres with radii d_{ij} centred in x_i and x_j, as we would do for an *RNG*, we look for points inside the hypersphere with diameter equal to $d(x_i, x_j)$ and centred in the midpoint between x_i and x_j.

Matula and Sokal (1980) called the *GG* the *least squares adjacency graph*, studied some of its properties in the context of pattern recognition, and obtained some interesting results, but only for 2 dimensional observations.

It should be noted that the above definition of *GG* implies that the observations are in an Euclidean space, unlike those for the *RNG*, the *MST* or the *NNG*. However, it is possible to obtain *GG* by applying least squares adjacency criterion to the data distance matrix.

It is straightforward to use Lefkovitch's ideas for *RNG*s to obtain generalized *GG*s. The method summarized in equation (1.4) can be used for this purpose. An example showing the first two *GG*s for 50 points in the plane appears in Figure 1-13.

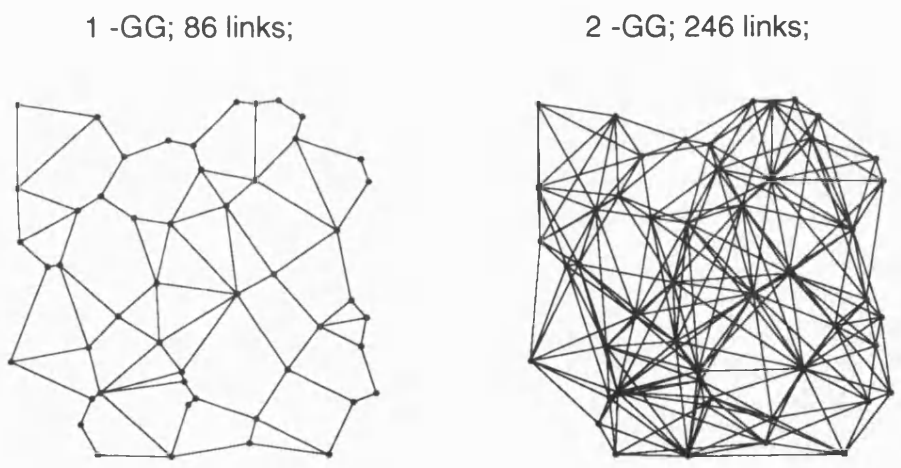


Figure 1-13: Generalized Gabriel Graphs

Urquhart's (1980, 1982) algorithm usually runs in much less than $\mathcal{O}(N^3)$ time. In order to assess its behaviour we generated one hundred graphs over a set of N uniformly distributed points in p dimensions. Table 1.1 presents the average number of operations needed to calculate an *RNG* and a *GG* using Urquhart's algorithm.

N	p	1-*RNG*	1-*GG*
10	2	133.54	165.27
10	3	135.23	194.03
10	5	137.35	241.49
10	10	140.63	304.19
10	20	142.15	349.08
20	2	756.67	987.19
20	3	758.29	1220.66
20	5	788.28	1673.16
20	10	823.94	2511.66
20	20	846.02	3177.15
50	2	6031.28	8330.20
50	3	6427.96	11591.90
50	5	6930.73	18476.50
50	10	7201.66	33607.30
50	20	7564.91	50823.00
100	2	27634.93	38720.91
100	3	29771.20	56974.06
100	5	32633.21	100266.91
100	10	36532.80	221026.10
100	20	37992.21	386974.04

Table 1.1: Urquhart's algorithm performance: *RNG* and *GG*

It appears that for the *GG*, increasing the dimension of the data points has a detrimental effect on the performance of the algorithm. For the *RNG* there seems to be an increase in the efficiency of the algorithm for higher dimensional configurations. For *GG* Urquhart's algorithm is far from being efficient; however, it is still feasible to use it for moderately large sample sizes.

The complexity of the *GG* increases with the dimensionality of the data much faster than it does for the *RNG*. Indeed, we observed that for higher dimensions, and for moderately large number of points, it almost coincides with the complete graph. Thus, in that case, it is not uncommon to have very few differences in the edges of 1-*GG* and

the 2-*GG*.

1.8 Families of Limited Neighbourhood Graphs

The *RNG* and *GG* can be expressed in terms of a *region of influence*, $\mathcal{R}$. In order to have two points linked in an associated graph $\mathcal{S}$, the $\mathcal{R}$ corresponding to that pair of points must be empty. For instance the $\mathcal{R}$ for the *GG* is a hypersphere; for the *RNG* it is a "lune". Urquhart (1982) proposed the following generalization to obtain families of graphs $\mathcal{S}_l$ based on influence regions $\mathcal{R}_l$:
two points x_i and x_j are linked in $\mathcal{S}_l$ iff

$$x_k \notin \mathcal{R}_l(x_i, x_j),\ \forall k = 1, \ldots, n, k \neq i \text{ and } k \neq j \tag{1.5}$$

where the $\mathcal{R}_l$ defining $\mathcal{S}_l$ can be written as:

$$\mathcal{R}_l(x_i, x_j) = \{x | f\,[d(x, x_i),\ d(x, x_j)] < d(x_i, x_j), i \neq j\} \tag{1.6}$$

where

$$d(x_i, x_j) = 0 \quad \Leftarrow \quad \mathcal{R}_l(x_i, x_j) = \varnothing \tag{1.7}$$

and

$$\mathcal{R}_l(x_i, x_j) = \mathcal{R}_l(x_j, x_i). \tag{1.8}$$

i.e. f is well behaved in the sense of yielding a finite nonempty region for $d(x_i, x_j) > 0$. The *RNG* and the *GG* can be defined in this way given the following regions of influence:

$$\mathcal{R}_{RNG}(x_i, x_j) = \{x \mid \max[d(x, x_i),\ d(x, x_j)] < d(x_i, x_j),\ i \neq j\} \tag{1.9}$$

and

$$\mathcal{R}_{GG}(x_i, x_j) = \{x \mid d^2(x, x_i) + d^2(x, x_j) \leq d^2(x_i, x_j),\ i \neq j\} \tag{1.10}$$

The regions described in equations (1.9) and (1.10) are referred as the *lune* and the *disc* for points x_i, x_j (Figure 1-14).

Figure 1-14: $\mathcal{R}_{RNG}$ $\mathcal{R}_{GG}$

Urquhart (1982) proposed three specific families of limited neighbourhood regions:

$$\mathcal{R}_1(x_i, x_j, \sigma) = \mathcal{R}_{RNG} \cup \{x \mid \sigma \min[d(x, x_i), d(x, x_j)] < d(x_i, x_j)\} \tag{1.11}$$

$$\mathcal{R}_2(x_i, x_j, \sigma) = \mathcal{R}_{GG} \cup \{x \mid \sigma \min[d(x, x_i), d(x, x_j)] < d(x_i, x_j)\} \tag{1.12}$$

$$\mathcal{R}_3(x_i, x_j, \sigma) = \{x \mid [d^2(x, x_i) + d^2(x, x_j)](1 + \sigma) < d^2(x_i, x_j)\} \tag{1.13}$$

Examples of 2-d regions of these families appear in Figure 1-15. The third region corresponds to concentric hyperspheres with radii

$$\frac{1-\sigma}{1+\sigma} d^2(x_i, x_j)$$

and centred at the midpoint along the line connecting x_i and x_j.

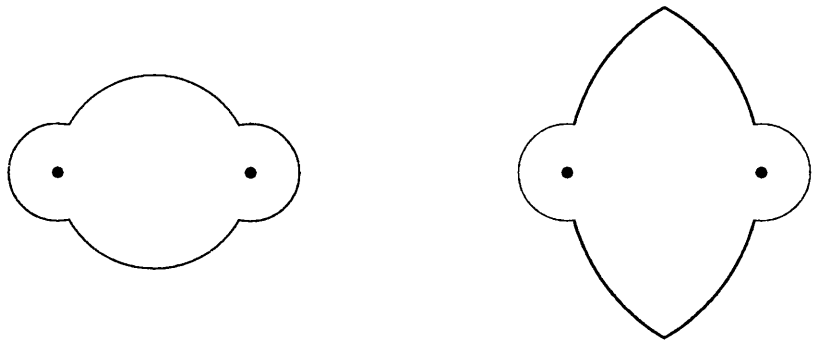


Figure 1-15: $\mathcal{R}_1(0.25)$ $\mathcal{R}_2(0.25)$

It is interesting to note that the the *mutual nearest neighbour graph* of order 2, $\mathcal{S}_{MNN}(2) \in \mathbf{S}$ is generated from the region of influence:

$$\mathcal{R}_{MNN}(x_i, x_j, 2) = \{x \mid \min[d(x, x_i), d(x, x_j)] < d(x_i, x_j)\} \tag{1.14}$$

This region produces graphs with very few edges, as only points which are mutual nearest neighbours will be linked (Figure 1-16). Clearly,

$$\mathcal{R}_{MNN}(x_i, x_j, 2) = \mathcal{R}_1(x_i, x_j, 1) = \mathcal{R}_2(x_i, x_j, 1)$$

Another family may be defined as:

$$\mathcal{R}_4(x_i, x_j, \varepsilon) = \{x \mid [d(x, x_i) + d(x, x_j)]\,\varepsilon < d(x_i, x_j)\} \tag{1.15}$$

where $0 < \varepsilon < 1$ is a parameter analogous to the eccentricity of an ellipse with foci in x_i and x_j

Cluster analysis methods are a natural field of application for these families of graphs. By associating with each edge of the *RNG* or the *GG* a dissimilarity $d^* = 1/\sigma^*$, σ^* being the value of the parameter which causes the edge to be deleted, it is possible

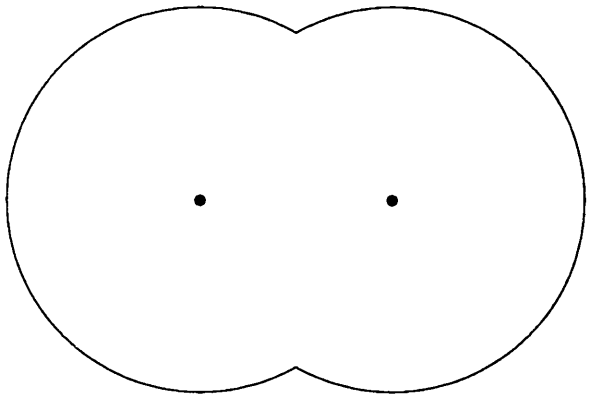


Figure 1-16: $\mathcal{R}_{MNN}$

to define dendrograms based on these families of graphs. As these graphs usually produce a disconnected graph $\mathcal{S}_l$, the connected subgraphs within it induce a partition of the observations. Urquhart (1982) has shown that the resulting clusters will satisfy several consistency and stability criteria that make them an interesting alternative to other clustering procedures. Some examples appear in Figures 1-17 and 1-18.

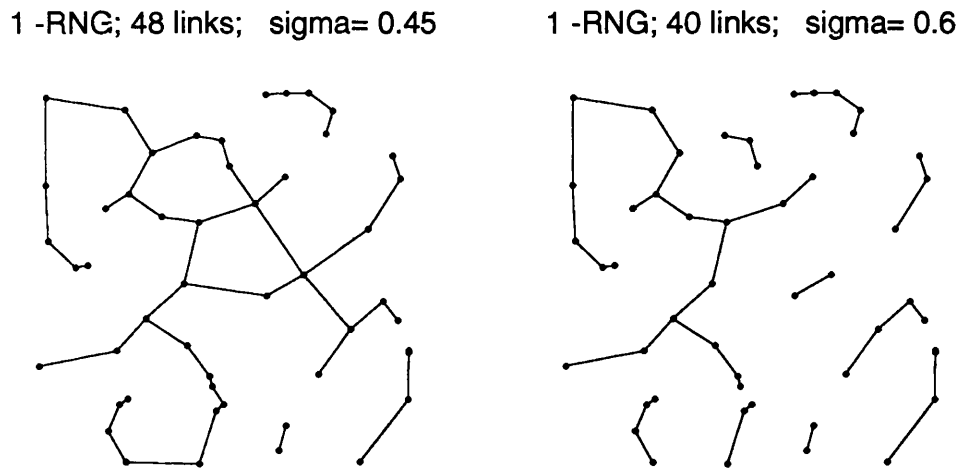


Figure 1-17: $\mathcal{S}_1$

The following lemma, due to Urquhart (1982), gives an interesting property for the *RNG*.

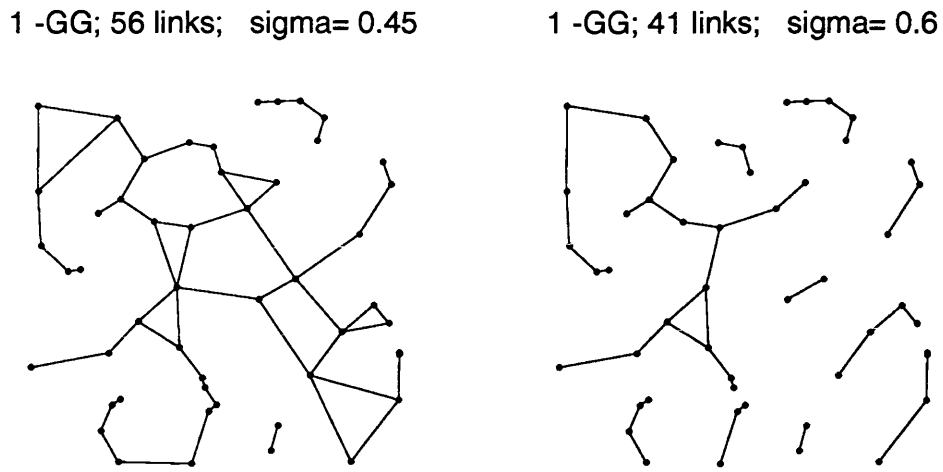


Figure 1-18: $\mathcal{S}_2$

Lemma 1 *$\mathcal{R}_{RNG}$ is the maximal region of influence $\mathcal{R}_I \in \mathcal{R}$ that is guaranteed to give a connected graph $\mathcal{S}_I$.*

Proof: From equation (1.6), we have that a necessary condition for $\mathcal{S}_I \in \mathcal{S}$ to be connected is that every point is linked to its nearest neighbour; so, $\mathcal{S}_I$ will not necessarily be connected. Now, the *RNG* is the largest region that guarantees links between nearest neighbours.

A consequence of this lemma is that any graph $\mathcal{S}_I$, where $\mathcal{R}_I \cap \overline{\mathcal{R}}_{RNG} \neq \emptyset$ may be disconnected, and thus such graphs may be useful for detecting clusters as well as points from different samples which are close together.

Indeed, these graphs might not be spanning graphs, i.e., some points can appear isolated in the graph. This is not an advantage in the context of hypothesis testing, although it could be possible to construct a family of graphs by modifying Urquhart's definitions in order to have the 1-*NNG* as the most disconnected graph allowed.

A method for obtaining $\mathcal{S}_1$ and $\mathcal{S}_2$ is to delete the edges (x_i, x_j) of the *RNG* or the *GG* if the ratio of $d(x_i, x_j)$ to min $[d(x_i, x_a), d(x_j, x_b)]$ is greater than σ, where x_a and x_b denote the nearest neighbours in the *RNG* or the *GG* to x_i and x_j, respectively, and $x_a \neq x_j, x_b \neq x_i$.

If we also preserve in $\mathcal{S}_1$ and $\mathcal{S}_2$ those edges linking x_a, x_i, x_b, x_j, we obtain graphs whose changes with respect to the parameter σ are described in the following table.

σ	$\mathcal{S}_1$	$\mathcal{S}_2$
1	1–*MNNG*	1–*MNNG*
0	*RNG*	*GG*

The families of graphs resulting from $\mathcal{R}_3$ and $\mathcal{R}_4$ have the following behaviour:

$\mathcal{S}_3(1) =$	$\mathcal{K}_N$	$\mathcal{S}_4(1) =$	$\mathcal{K}_N$
$\mathcal{S}_3(0) =$	GG	$\mathcal{S}_4(0) =$	$\mathcal{E}_N$
$\mathcal{S}_3(-1) =$	$\mathcal{E}_N$		

Urquhart's families of graphs are useful tools to define clusters. However, there is an element of arbitrariness in the selection of the parameter σ, and thus, we do not use them to define test statistics. The insight that they provide as a framework for the *RNG* and the *GG* justifies their discussion in this chapter.

1.9 Delaunay Triangulations

A *triangulation* of a metric space is a planar subdivision in which all its bounded regions are simplexes. A *triangulation of a finite set S* of points is a planar graph on S with the maximum number of edges. This is equivalent to saying that the triangulation of S is obtained by joining its points by nonintersecting straight line segments so that every region internal to the convex hull of S is a simplex. Suppose we have N distinct data points with positions $x_1, \ldots, x_N$ in a metric space. If we assign to each data point the territory that is nearer to it than to any other data point, we will induce a partition of the whole space. If the N points, denoted by $\mathcal{X}$, are in the plane, let T_i denote the subset of $\mathbb{R}^k$ which contains all the points closer to x_i than to any other element of $\mathcal{X}$. Clearly, T_i is an the open convex polygon and can be expressed as $\bigcap_{j \neq i} H_{ij}$, where H_{ij} is the open half plane containing x_i bounded by the perpendicular bisector of x_i and x_j. T_i

is the *Voronoi polygon* corresponding to x_i. The *Dirichlet tessellation* is the collection of these polygons. The Delaunay triangulation, *DT*, is the graph obtained by joining the points which share a side of their Voronoi polygons with length greater than 0.

For p dimensional Euclidean space, the Delaunay triangles are simplexes with $p+1$ data point as vertices. Each vertex in the Dirichlet tessellation is the point where $p+1$ territories meet and is the centre of the hypersphere passing through all the vertices of the associated simplex. In two dimensions the vertices of the tiles occur where three territorial boundaries meet; also, each of these vertices is equidistant from the three data points associated with the territories defining it. There are degenerate cases for this condition. For instance, in the plane, if the points are positioned in a regular square lattice, then the vertices correspond to points where four boundaries meet. In higher dimensions there are many more possibilities for having degenerate points in a *DT*.

Preparata and Shamos (1985) mentioned the following theorem for planar configurations.

Theorem 4 $GG \subseteq DT$

Proof: Let $\overline{x_i x_j} \in GG$ and let $\mathcal{D}$ be the interior of the sphere with diameter $d(x_i, x_j)$ and $\mathcal{E}$ its boundary. Then there is no other point of the pattern inside $\mathcal{D}$. To see that this implies that $\overline{x_i x_j} \in DT$ consider two any other points x_k and x_l. If both of them are on $\mathcal{E}$, then, no matter how close x_k and x_l are to, say, x_i, there will always be a side of non-zero length between x_i and x_j in the tessellation. Furthermore, if at least one of the other points lies in $\bar{\mathcal{D}} = (\mathcal{D} \cup \mathcal{E})^c$, the length of this common side will be even greater Figure 1-19 illustrates this point. The dotted lines in that Fig. represent the boundaries of the tiles defined by the Delaunay triangulation.

The *DT* can be constructed from any distance matrix. However, the vast majority of its applications have been to 2 dimensional data sets using Euclidean distance. Green and Sibson (1978) presented an algorithm to construct the *DT* of a set of points in the plane which runs in $\mathcal{O}(N \log N)$ time. Sibson (1980) reviewed some applications of the *DT* in data analysis. Boots and Murdoch (1983) gave an extense bibliography on applications of this graph to several subjects. Hinde and Miles (1980), and Quine and Watson (1984) have presented detailed simulation studies of the distribution of several

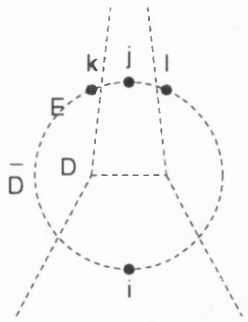


Figure 1-19: $GG \subseteq DT$

attributes of *DT*s in the plane. They have provided approximations to the distributions of the number of sides, perimeter, area and inner angles of the polygons generated by the *DT* with respect to a planar Poisson process.

Bowyer (1981) and Watson (1981) gave algorithms for computing *DT*s for p-dimensional points. Their algorithms run in $\mathcal{O}(a_p N^{(1+1/p)} + b_p N)$ and $\mathcal{O}(N^{(2p-1)/p})$ times, respectively. However, for Bowyer's algorithm, its author mentioned that the coefficient b_p grows very quickly as p increases.

DT; links= 123

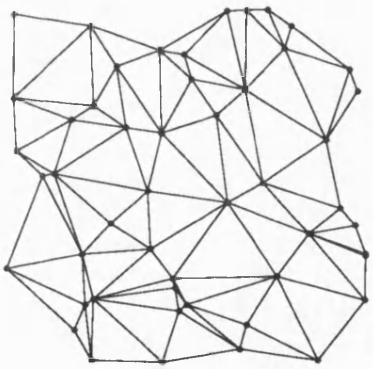


Figure 1-20: Delaunay Triangulation

Both algorithms are difficult to use and can be computationally very expensive for high dimensional data. Bowyer's algorithm requires the specification of a p-dimensional simplex enclosing the sample points in order to begin to compute the *DT*. This

introduces an element of arbitrariness in the resulting graph, as some links between points near the edge of the convex hull of the observations may change according to the initial simplex. Figure 1-20 was produced using the package TILE4 of the University of Bath (Sibson (1981)). A dual problem for the *DT* is the construction of the *p*-dimensional convex hull for a set of points. A general algorithm to solve this problem is still an open question in computational geometry. Jóźik (1983) outlined an algorithm to find multivariate convex hulls, but imposing several geometrical conditions on the data.

Figure 1-21 shows the average number of links for selected graphs based on applying Euclidean distance to 100 configurations of uniformly distributed points in 2 and 20 dimensions.

For every combination of the number of nodes and of dimensions considered, the observed number of spanning trees is always larger than the expected number (equations (1.2) and (1.3)). This does not happen for the *n-MST* or the *n-NNG*. Furthermore, the *RNG* has the largest ratios between observed and expected complexities of all the graphs, suggesting a more efficient selection of the edges. Although it has been proved that 1-*MST* $\subseteq$ 1-*RNG*, both graphs very seldom coincide. This happens only for configurations with about 10 points or less. The proportion of coincidences observed for these configurations decreases with the dimensionality of the data. These facts were observed by Lefkovitch (1984).

As our interests are in general procedures, and in view of the difficulties encountered to calculate the *DT* for higher dimensions, we did not use this graph to construct tests. As in the case of the Urquhart's graphs, the insight gained by briefly reviewing the *DT* justifies its inclusion in this chapter.

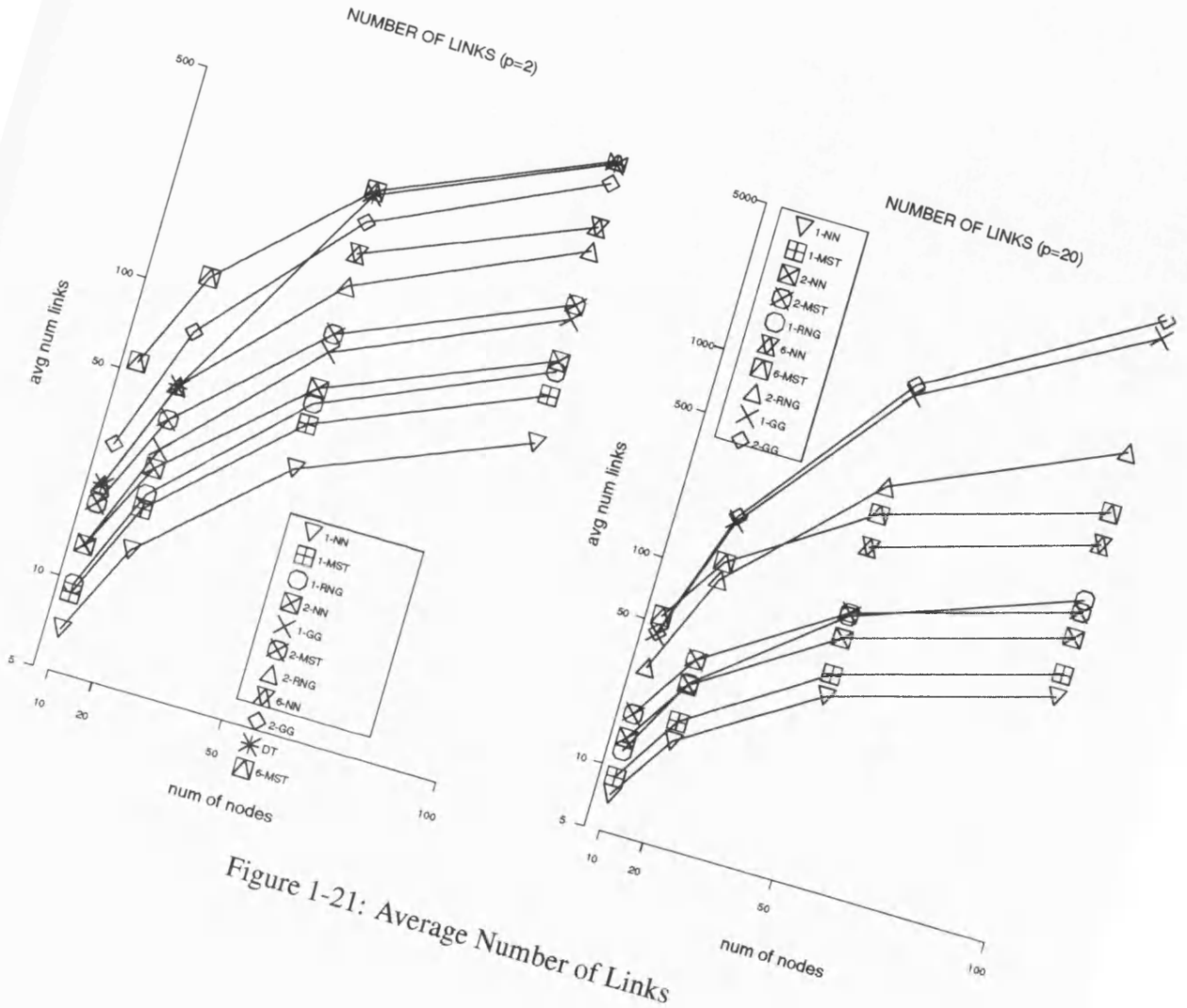


Figure 1-21: Average Number of Links

Chapter 2

Generalizations of the Multivariate Runs Test

2.1 Introduction

The first kind of multivariate tests to be discussed is based on the Friedman and Rafsky (1979) version of the Wald-Wolfowitz (1940) runs test. In this chapter we also examine some aspects of the theory of generalized correlation coefficients (*GCC*). We do so as the notation and some results from *GCC*s will be used later on. We begin by presenting a brief discussion of the nonparametric approach to hypothesis testing for multivariate data.

Methods based on ranks provide alternative procedures to the classical parametric approach to test the null hypothesis of homogeneity of K populations

$$H_0 : F_{X_1} = F_{X_2} = \cdots = F_{X_K}. \tag{2.1}$$

Rank tests have null distributions based on the permutations of the ranks over the sample values. These tests do not require one to assume that the distribution of the data belongs to some family of distribution functions specified by a finite number of parameters. Rank tests are usually constructed by conditioning on the minimal

sufficient statistic, that is, by regarding the order statistics as fixed and using the consequences that under the null hypothesis all permutations of the ordered values are equally likely (Cox and Hinkley, 1974).

Many univariate nonparametric K-sample tests for general alternatives are based on statistics computed over the ranks induced on the observations by sorting the pooled data. For nonparametric multivariate statistics, a common choice is to condition on the ranks calculated separately for each variable. This method has been extensively studied in the book by Puri and Sen (1971).

There are not many nonparametric multivariate tests which are conceptually different from those encompassed by Puri and Sen. As an example, of a different approach, Chung and Fraser (1958) presented a randomization procedure for a two-sample multivariate test, which does not seem likely to be generalized for the K-sample homogeneity problem.

The nonparametric multivariate tests we are interested in do not follow Puri and Sen's or Chung and Fraser's approaches. Instead of working with the ranks obtained for each individual variable, we condition on the relationships generated between pairs of observations by means of some graph constructed without any reference to the sample identity of the data. This is analogous to conditioning on the ranks of the sorted pooled data, as we might expect that the graph employed conveys the relationships of nearness which would be reflected by the ranks of the pooled observations in the univariate case. Following this approach, we obtain distribution free statistics.

We are interested in testing the null hypothesis of homogeneity for K populations against the class of alternatives that violate it. In the multivariate case, it is important to provide a distribution free alternative to parametric procedures, as these are usually based on the assumption of multivariate normality for the particular data. Testing the goodness-of-fit for normality, or other particular conditions, for example, the equality of variance-covariance matrices, can be difficult.

If it is found that the assumptions do not hold, or if it is not possible to evaluate to what extent they do. So, there are three possibilities:

1. To assume that the parametric procedure is robust enough to handle violations

of the assumptions. If this is not the case, the significance level of the test may be severely affected.

2. To transform the data in order that they may fit the method's assumptions.

3. To use a nonparametric procedure.

We advocate the use of the third way of proceeding.
Following Scholz and Stephens (1987) we distinguish three important features involved in methods for testing the K-sample homogeneity hypothesis against general alternatives.

1. They are useful aids for establishing differences in several sampled populations with particular sensitivity towards the extremes of the pooled sample.

2. They are a natural way for judging if several samples are homogeneous enough in order to be pooled together for further analysis.

3. They can be more effective than these methods designed to be consistent only against a rather restricted set of alternatives; this is of chief importance for multivariate data, as the precise characterizations of alternatives can be a difficult task to achieve.

The Friedman-Rafsky (1979) multivariate runs two-sample test uses the number of edges linking points from different samples on the minimal spanning tree of the pooled sample as a proper analogue for the number of runs in a univariate sequence.
In another paper, four years after their first one, these authors outlined a slight modification to this test: working with the number of edges on an interpoint distance graph which links points with the same sample identity, they proposed an equivalent statistic to the one used in the multivariate runs test. In the next sections, we specify the class of hypotheses we are interested in and some alternatives to them. Later on in this chapter, we discuss a multivariate K-sample generalization of the runs test. This is done in the context of statistics based on the intersection of graphs written as *GCCs*. Such a procedure proved to be an efficient way of presenting the calculations needed

to obtain the higher moments of the test statistic, so we start with a description of these statistics.

2.2 Hypotheses Specification

In this section we establish the framework that will be used in further sections, and discuss the null hypothesis and the alternatives of interest. In doing so, we follow the notation and conventions used by Puri and Sen (1971, §5).

We assume that we observe individuals from K p-valued populations. The j-th sample size will be denoted by n_j, and the total sample size by $N = \sum_{j=1}^{K} n_j$.

Let

$$\{\mathbf{X}_\alpha^{(j)} = (X_{1\alpha}^{(j)}, \ldots, X_{p\alpha}^{(j)})'\},$$

where $\alpha = 1, \ldots, n_j, j = 1, \ldots, K$, be a set of independent multivariate random values. The cumulative distribution function (*cdf*) is denoted by $F_{X_j}(\mathbf{x})$. It is assumed that each of these *cdf*s belongs to some class of distributions functions C.

We now assume that C is the class of all continuous distribution functions. This is done mainly in order to keep the notation consistent with that used by Puri and Sen. In actual fact, the continuity assumption is not at all essential for the generalized runs tests, as the only requirement is that a distance measure can be obtained from the data.

The mean vector of the j-th population will be denoted by

$$\mu^{(j)} = \begin{pmatrix} \mu_1^{(j)} \\ \vdots \\ \mu_p^{(j)} \end{pmatrix}$$

and its $(p \times p)$ covariance matrix by $\Sigma^{(j)}$. If the p components of $\mathbf{X}_\alpha^{(j)}$ are uncorrelated

for $\alpha = 1, \ldots n_j$, then

$$\Sigma^{(j)} = \begin{pmatrix} \sigma_1^{(j)} & & \\ & \vdots & \\ & & \sigma_p^{(j)} \end{pmatrix} \cdot \mathbf{I}_p$$

where $\mathbf{I}_p$ identity matrix of order p.

The hypothesis of homogeneity to be tested, say H_0, can be written as:

$$H_0 : F_{X_1}(\mathbf{x}) = \ldots = F_{X_K}(\mathbf{x}) = F(\mathbf{x}) \tag{2.2}$$

for all $\mathbf{x}$, for some $F \in \mathcal{C}$.

The alternative to H_0 is the hypothesis that specifies that (2.2) does not hold. Two types of hypotheses contained in this class are the *location shift* and the *scale* alternatives. For the first type, let

$$H_1^{(t)} : F_{X_j}(\mathbf{x}) = F(\mathbf{x} + \delta_j), \qquad \text{for all } j = 1, \ldots K \text{ and for } F \in \mathcal{C}; \tag{2.3}$$

then, the hypothesis of homogeneity can be written as:

$$H_0^{(t)} : \delta^{(1)} = \ldots = \delta^{(K)} = \mathbf{0} \qquad \equiv \qquad \mu^{(1)} = \ldots = \mu^{(K)} \tag{2.4}$$

against the alternatives that $\delta^{(1)}, \ldots, \delta^{(K)}$ are not all equal.

For the latter type, if

$$F_{X_j}(\mathbf{x}) = F_{\mathbf{X}_j^*}(\mathbf{x}), \quad \text{with } \mathbf{X}_j^* = \left(\frac{x_1 - \mu_1}{\sigma_1^{(j)}}, \cdots, \frac{x_p - \mu_p}{\sigma_p^{(j)}} \right), \tag{2.5}$$

and $\sigma^{(j)} = (\sigma_1^{(j)}, \ldots, \sigma_p^{(j)})$, for all $j = 1, \ldots, K$, then the null hypothesis corresponds to

$$H_0^{(s)} : \sigma^{(1)} = \ldots = \sigma^{(K)} = \mathbf{1} \qquad \equiv \qquad \Sigma^{(1)} = \ldots = \Sigma^{(K)} \tag{2.6}$$

against the alternatives that $\sigma^{(1)}, \ldots, \sigma^{(K)}$ are not all equal. It should be noted that the scale type alternatives assume the homogeneity of the location vectors of the distributions $F_{X_1}, \ldots, F_{X_K}$.

These ideas apply if the data are measured with a scale at least rational. For ordinal and nominal data, the runs tests would still work, although the alternative hypothesis of interest would be the heterogeneity of the populations.

2.3 Generalized Correlation Coefficients

In many nonparametric methods, the test statistic can be expressed in terms of the number of common edges shared by two graphs, each of them containing relevant information about the neighbourhood relationships existing within two different sets of variables.

We now discuss some concepts concerning this kind of statistics. The easiest way of doing so is within the framework of generalized correlation coefficients. The theory of *GCC* was first presented by Daniels (1944). A comprehensive reference is the book by Kendall (1962).

Consider a sample $(x_i, y_i), i = 1, \ldots, N$ of ordered pairs, and let a_{ij} and b_{ij} be scores for every pair (i, j) of X and Y observations respectively. Then, up to some form of standarization, a *GCC* has the form

$$\Gamma = \sum_{i}^{N} \sum_{j}^{N} a_{ij}\, b_{ij} \tag{2.7}$$

As it is well known, for a suitable choice of the a_{ij} s and b_{ij} s, Pearson's r, Spearman's ρ, and Kendall's τ correlation coefficients may be expressed as Γ. If we condition on the observed values of X and Y, it is possible to test the null hypothesis of no correlation by ranking the observed value of Γ within the distribution of

$$\Gamma(\pi) = \sum_{i}^{N} \sum_{j}^{N} a_{ij}\, b_{\pi(i)\pi(j)} \tag{2.8}$$

where π is a permutation of the integers $(1, ..., N)$. This is an adequate procedure, since under the hypothesis of no correlation between X and Y all the permutations (i.e. X, Y pairings) are equally likely. This permutational distribution determines if the observed value of Γ is significant: too large or too small values of it would be evidence against the null hypothesis.

To test the hypothesis (2.1), no correlation refers to relationships between closeness in the multivariate space (X) and sample identity (Y): if they are positively correlated, then there should be evidence against the null hypothesis.

Although the scores a_{ij}, b_{ij}, and thus Γ, can be defined for pairs of multivariate observations (x_i, y_i), it is not possible to give straightforward generalizations for the notion of ordering used to define nonparametric correlation coefficients like ρ and τ. To overcome this difficulty, Friedman and Rafsky (1983) suggested the use of a *GCC* depending on the intersection of interpoint distance graphs.

Let $\mathcal{G}_X$ and $\mathcal{G}_Y$ be graphs defined over the X and the Y observations, respectively. The test statistic Γ_R is the number of edges in the intersection of the two graphs. Clearly, the value of this statistic will tend to be large if observations which are close in X also happen to be close in Y.

Let

$$a_{ij} = \begin{cases} 1 & \text{edge } (i,j) \in \mathcal{G}_X \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

and

$$b_{ij} = \begin{cases} 1 & \text{if edge } (i,j) \in \mathcal{G}_Y \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

then the statistic

$$\Gamma_R = \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} a_{ij}\, b_{ij} \tag{2.11}$$

is the number of edges in the intersection of $\mathcal{G}_X$ and $\mathcal{G}_Y$.

The set up introduced in this section will be used later on in connection with multivariate nonparametric association and prediction measures. In the following sections of this chapter we obtain the moments of Γ_R and discuss its null distribution. In the next two sections we present the Wald-Wolfowitz runs test and a K-sample generalization of the Friedman-Rafsky multivariate version of the two-sample runs test.

2.4 Wald-Wolfowitz Runs Test

Let F_{X_1} and F_{X_2} be two distribution functions. The null hypothesis can be specified as $H_0 : F_{X_1} = F_{X_2}$; the general alternative hypothesis is $H_1 : F_{X_1} \neq F_{X_2}$. The total sample size is $N = n_1 + n_2$. The Wald-Wolfowitz (1940) runs test requires sorting the pooled observations in order to count the number of groups of individuals from the same sample which appear contiguously in the sorted list; each of this groups is called a *run*. The total number or runs, R, is the test statistic. H_0 is rejected for small values of R, as it should indicate that the two samples are well separated. It is possible to obtain the exact permutational distribution of R. Under H_0, all the permutations of the sample identities over the pooled data have the same probability of occurring. The permutational distribution function can be written as:

$$\Pr(R=z) = \begin{cases} \dfrac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{N}{n_1}} & \text{if} \quad z=2k \\ \dfrac{\binom{n_1-1}{k}\binom{n_1-1}{k-1}+\binom{n_1-1}{k-1}\binom{n_1-1}{k}}{\binom{N}{n_1}} & \text{if} \quad z=2k+1 \end{cases} \tag{2.12}$$

for $z = 2, 3, \ldots, N$ and $k \in \mathbb{Z}^+$. To find the critical value for a type I error α, one has to obtain the integer z_0 such that

$$\sum_{z=2}^{z_0} \Pr[R=z] = \alpha$$

as nearly as possible.

In addition, the quantity

$$W = \frac{R - \dfrac{2n_1n_2}{N} - 1}{\left(\dfrac{2n_1n_2(2n_1n_2-N)}{N^2(N-1)}\right)^{\frac{1}{2}}}$$

has, asymptotically, a standard normal distribution. It is well known that the normal approximation is remarkably accurate, even for quite small total sample sizes.

The test is consistent if the ratio n_1/n_2 is bounded away from 0 and ∞ when n_1, $n_2 \to \infty$. Mood (1940) was the first to generalize the distribution (2.12) to the K-sample case. The book by David and Barton (1960) discusses in detail many interesting generalizations of the concept of runs.

It has been noticed by Smith (1953), Blumenthal (1963) and Capon (1965), amongst others, that the univariate Wald-Wolfowitz runs test is not very powerful in comparison with other two-sample nonparametric tests. However, Friedman and Rafsky (1979) found that their multidimensional generalization has good power in some cases and that it can be increased by a proper selection of the graph used to define the scores a_{ij}

and b_{ij} of equations (2.9) and (2.10). This has been confirmed by a modification of the Friedman-Rafsky test proposed by Whaley and Quade (1985). The difference with the original multivariate runs test is that the latter authors linked all pairs of observations whose distance is less or equal than a given threshold. Whaley and Quade (1985) remarked that better performances can be achieved by using orthogonal *MST*s instead of graphs based on threshold distances.

For the univariate case, Whaley (1987) presented a modification of the runs test based on threshold distances. His aim was to eliminate two possible drawbacks of the ordinary Wald-Wolfowitz test:

1. If there is an outlier, it has to be linked with another observation.

2. An observation can be linked only to one observation, even if it is very similar to more than one of them.

This author tried to find an optimal tolerance threshold distance in terms of the power of the two-sample test for shift alternatives. He also showed that the power of the univariate test does increase when observations are linked using these threshold-based links instead of the ones induced by simply connecting any point to its adjacent neighbours in the ordered list. However, his way of proceeding in order to find optimal threshold distances seems too complicated to be generalized.

2.5 K-sample Multivariate Runs Test

2.5.1 Multivariate Runs

We now present a K-sample version of the multivariate runs test. We discuss the exact and asymptotic distributions of the test statistics. Later in this chapter we study some approximations to the null distribution of Γ_R. To do so, we followed two approaches: one involves generating a sample from all the possible permutations of sample identities and obtaining the value of the statistic for each of them; in the other, we fit a Pearson distribution based on the values of the first four moments calculated under the null hypothesis.

There are some antecedents of multivariate nonparametric tests using interpoint distance graphs, but all of them have only addressed the two-sample problem.

Schilling (1986) proposed a two-sample test for multivariate observations based on the n-nearest neighbours graph. He considered only Euclidean distances in p-dimensional spaces. His results concerning the power of this test, together with those of Friedman and Rafsky (1979) indicate that it is possible to achieve good power for some alternatives using the two-sample multivariate generalization of the runs test or some similar test.

Henze (1988) generalized Schilling's results for the two-sample problem, using weighted proportions of nearest neighbours based on any distance measure in $\mathbb{R}^p$.

We extend Friedman and Rafksy's ideas for K samples using *MST*s and *NNG*s and also other graphs which have been described in the previous chapter.

To test the null hypothesis (2.1) of homogeneity of K populations, we can think of $\mathcal{G}_X$ as a spanning graph constructed from a distance matrix defined over the pooled observations X and of $\mathcal{G}_Y$ as

$$\mathcal{G}_Y = \bigcup_{j=1}^{K} \mathcal{K}_{n_j} \tag{2.13}$$

where $\mathcal{K}_{n_j}$ is the complete graph formed by linking all the observations in the j-th sample; i.e. its edges are defined by nodes from the same sample. This allows us to write the test statistic as a generalized correlation coefficient between points linked in $\mathcal{G}_X$ (which should be points close together in the p-dimensional space) and sample identity. To do so, we define scores a_{ij} to be equal to 1 if the nodes i and j form an edge in $\mathcal{G}_X$ and to be 0 otherwise. The scores b_{ij} are equally defined for the edges of $\mathcal{G}_Y$.

Suppose we have observations from K samples with sample sizes $n_1, n_2, \cdots, n_K$ from distributions $F_{X_1}, F_{X_2}, \ldots, F_{X_K}$. The pooled sample size is $N = \sum_{j=1}^{K} n_j$.

Define the *r.v.* Z_i to be

$$Z_i = \begin{cases} 1 & \text{if } i \in \mathcal{G}_Y, \qquad 1 \le i \le e_X \\ 0 & \text{otherwise} \end{cases} . \tag{2.14}$$

Equation (2.11) may be written as:

$$\Gamma_R = \sum_{i=1}^{e_X} Z_i$$

where e_X is the number of edges of $\mathcal{G}_X$. Γ_R is the number of edges which are common to both graphs. Under H_0, we should not observe a close correspondence between the nodes linked in $\mathcal{G}_X$ and those linked in $\mathcal{G}_Y$. Thus the values of Γ_R that would lead us to reject H_0 will be relatively large. The multivariate runs test statistic is defined to be 1 plus the number of edges on an interpoint distance graph which link observations from different samples. This is a direct analogy to the univariate runs statistic. The original Friedman-Rafsky multivariate runs test statistic R can be seen as based on a graph $\mathcal{G}_Y'$ whose edges correspond to pairs of nodes from different samples. The value of R is equal to 1 plus the number of edges in $\mathcal{G}_X$ that link nodes from different samples. Rejection of H_0 is indicated by observing relatively small values of R. In fact, the relation between Γ_R and R is expressed as:

$$R = e_X - \Gamma_R + 1.$$

We chose to work with Γ_R defined using $\mathcal{G}_Y$ rather than $\mathcal{G}_Y'$ as the algebra necessary for the higher moments of the statistic is much more easily handled with the former graph. As Whaley (1983) pointed out, other statistics may be expressed, as we did with Γ_R, in the form of a *GCC*. Amongst them we can mention Mantel (1967) and Mantel and Valand (1970) space-time clustering statistics, Cliff and Ord (1981) spatial autocorrelation index $\mathcal{I}$ and some cases of Mielke et al. (1976) *MRPP* statistics.
Several approximations to the null distributions of this class of statistics have been proposed. Constanzo et al. (1983) have discussed higher moments approximation for a spatial autocorrelation index. Whaley (1985) worked out an approximation for his run

test based on a χ^2 distribution fitted using the first three moments. Semyaticki (1978), and Cliff and Ord (1981) obtained four-moments approximations to the distributions of the Mantel and Valand statistic, and to $\mathcal{I}$, respectively. Tracy and Tajuddin (1985), and Mielke, Berry and Wong (1986) did the same for *MRPP*-type statistics. All those authors mentioned that normal approximation should not be automatically taken for granted. Actually, for some of these statistics, the sample sizes required for the latter approximation to work satisfactorily seem to be rather large.

In the following subsection we discuss the limiting distribution of Γ_R. There we apply the arguments developed by Friedman and Rafsky for the multivariate runs test statistic, which can be used to prove that Γ_R has, asymptotically, a normal distribution.

2.5.2 Limiting Distribution of Γ_R

In view of well known results about the fast convergence to the normal distribution for the univariate runs test statistic, it would be plausible to expect that the null distribution of Γ_R should tend to the normal distribution fairly quickly. In order to have some form of evaluating how quickly is the convergence to the limiting distribution achieved, as well as having a way of fitting more exact approximations, we obtain the first four moments of the sampling distribution of Γ_R in the next subsection.

We now present Friedman and Rafksy's arguments to show that the permutational distribution of Γ_R approaches the normal distribution as $N \to \infty$, provided that some assumptions about the degree sequence of the observed graph and about the sample sizes are fulfilled. These ideas are based on the main result of Daniels (1944) which says that, under very mild conditions, the limiting distribution of *GCC*-type statistics is normal.

Daniels' result depends on the conditions

$$\sum_{i,j,k}^{N} a_{ij}\, a_{ik} \approx N^3 \qquad \text{and} \qquad \sum_{i,j,k}^{N} b_{ij}\, b_{ik} \approx N^3 \tag{2.15}$$

which may be replaced by the weaker conditions

$$\lim_{N\to\infty} \sum_{i,j,k,l} (a_{ij}\, a_{ik}\, a_{il})^2 \Big/ \sum_{i,j,k} (a_{ij}\, a_{ik})^3 = 0 \qquad (2.16)$$

with similar conditions for the scores b_{ij}.

In the context of *GCC* with scores defined as in equations (2.9) and (2.10), if d_i denotes the degree of the i-th node in any graph, conditions (2.15) and (2.16) can be written as

$$\sum_{i=1}^{N} d_i^2 \approx N^3 \qquad \text{as} \qquad N \to \infty \qquad (2.17)$$

and

$$\lim_{N\to\infty} \left(\sum_{i=1}^{N} d_i^3\right)^2 \Big/ \left(\sum_{i=1}^{N} d_i^2\right)^3 \to 0 \qquad (2.18)$$

respectively, for both $\mathcal{G}_X$ and $\mathcal{G}_Y$. These conditions put some restrictions on the topology of the graphs in order to insure a limiting normal distribution for Γ_R.

Condition (2.17) implies that the spanning subgraphs should be dense, i.e. they should contain a large proportion of the edges of the complete graph. To insure this, it is sufficient that the degree of each node grow linearly with N. As the sum of the degrees is twice the number of edges in any graph, and if e_X, e_Y and $e_{\mathcal{K}_N}$ denote the number of edges in $\mathcal{G}_X$, $\mathcal{G}_Y$ and $\mathcal{K}_N$, respectively, we have that, in order to insure asymptotic normality of Γ_R, e_X and e_Y must grow quadratically in N, or linearly in $e_{\mathcal{K}_N}$.

Even if the spanning graphs are sparse, condition (2.18) allows Daniels' results to hold. This is the case for n-orthogonal *MST*s and *n-NNG*, in a p-dimensional Euclidean space when n is fixed while N grows: both graphs are very sparse, having a maximum degree bounded by a constant independent of N, depending only on the dimension p. In this case, the expression in the l.h.s. of equation (2.18) is bounded by N^{-1}, thus assuring that Daniels' conditions hold.

As Friedman and Rafsky (1979) pointed out, in order to have a limiting normal distribution it is important that when N increases, the number of edges should remain distributed amongst the nodes in such a way that it avoids the situation of a too rapidly

decreasing proportion of the nodes defining a too rapidly increasing fraction of the edges.

It is possible to find sequences of sparse graphs which do not satisfy these conditions. For example, graphs like 'fan' trees with N nodes and having its edges defined in such a way that one node has always degree $N-1$ and the rest of the nodes have degree 1. On the other hand, if n_e is the number of nodes with degrees equal to $(N-l)$, with $l \geq 1$ and there are $(N-n_e)$ nodes with degree l, with N much larger than l, it is sufficient to have n_e growing linearly on N in order to achieve asymptotic normality, as in this case expression (2.18) is again bounded by N^{-1}.

Clearly, graphs for which the necessary conditions to attain an asymptotic normal distribution do not hold would not produce homogeneity tests with good power, as too many edges considered in the graph will basically give redundant information excluding a large proportion of edges that should highlight important features of the samples' situation. On the other hand, if we make $\mathcal{G}_X = \mathcal{K}_N$, then we would not be able to discriminate between useless and useful links according to the class of alternatives which we are interested to test against the homogeneity of the K populations. Graphs like orthogonal sequences of *MST*s or low-order generalized relative neighbourhood graphs provide a convenient balance between these two extremes.

All the lower order spanning graphs based on a distance matrix ($\mathcal{G}_X$) that were mentioned in the previous chapter satisfy condition (2.18), as they all are based on principles that forbid any node to have relatively large number of neighbours while a substantial proportion of them are isolated or have very few neighbours as the number of nodes grows.

Γ_R is constructed as the intersection of graphs defined by some neighbourhood relationships based on interpoint distances ($\mathcal{G}_X$) with graphs defined as the union of complete graphs defined within the K samples ($\mathcal{G}_Y$). To insure asymptotic normality of Γ_R, the scores based on this last graph have to satisfy the conditions discussed above. Basically, this means that the proportion of nodes from each sample has to be bounded away from 0 and 1 when N tends to infinity. This condition is a necessary one for getting a limiting normal distribution for several other nonparametric multisample tests (Puri

and Sen (1971)).

So we have seen that Γ_R has an asymptotic normal distribution under the null hypothesis of no correlation between the edges in the intersection of $\mathcal{G}_X$ and $\mathcal{G}_Y$. To have a better idea about the adequacy of the normal approximation, for smaller sample sizes, we now present the construction of the first four moments of Γ_R.

2.5.3 Moments of Γ_R

The moments of Γ_R can be calculated in a straightforward manner. The methodology used here is similar to those presented by Moran (1948), Barton and David (1966), Cliff and Ord (1981) and Friedman and Rafsky (1983).

The expected value of Γ_R offers no problems.

$$\mu_1(\Gamma_R) = \mathrm{E}(\Gamma_R \mid e_X, e_Y) = \mathrm{E}\left(\sum_{i=1}^{e_X} Z_i\right) = e_X \Pr[Z_s = 1] = e_X e_Y \binom{N}{2}^{-1} \qquad (2.19)$$

and e_X and e_Y can be expressed as:

$$e_X = \tfrac{1}{2}\sum_{i=1}^{N} \deg_X(v_i) \qquad e_Y = \sum_{j=1}^{K} \binom{n_j}{2} \qquad (2.20)$$

where v_i denotes the i-th node in the pooled sample, and $\deg_X(\cdot)$ denotes the degree of any node in $\mathcal{G}_X$.

Note that $\mathrm{E}(\Gamma_R)$ is independent of the topology of $\mathcal{G}_X$ and is conditioned only on the number of edges of each graph. This does not happen for the higher moments of the statistic: in general, the moment of order r can be expressed in terms of the K sample sizes and of the observed numbers of different subgraphs of $\mathcal{G}_X$ that can be formed using r edges.

The second central moment of Γ_R is obtained as:

$$\mu_2 = \mu_2' - \mu_1^2$$

$$= \mathrm{E}\left(\sum_{i=1}^{e_X} Z_i\right)^2 - \mu_1^2$$

$$= \sum_{i=1}^{e_X} \mathrm{E}(Z_i^2) + 2\sum_{i<j}^{e_X} \mathrm{E}(Z_i Z_j) - \mu_1^2$$

$$= \mu_1 + 2\sum_{i<j}^{e_X} \mathrm{E}(Z_i Z_j) - \mu_1^2 \qquad (2.21)$$

The value of $\mathrm{E}(Z_i Z_j)$ depends on whether or not the two edges i and j have a common node on a graph.

We need to consider two configurations:

$\mathcal{C}_1$ $\mathcal{C}_2$

Thus we have

$$\sum_{i<j}^{e_X} \mathrm{E}(Z_i Z_j) = p_1 \Pr[Z_s Z_t = 1 | \mathcal{C}_1] + p_2 \Pr[Z_s Z_t = 1 | \mathcal{C}_2] \qquad (2.22)$$

where p_m is the number of observed pair of edges appearing as in configuration $\mathcal{C}_m$ within the spanning graph $\mathcal{G}_X$, and s, t denote any two edges in $\mathcal{G}_Y$. These numbers can be expressed as

$$p_1 = \frac{1}{2}\sum_{i=1}^{N} \deg_X(v_i)\,(\deg_X(v_i) - 1) = C_X \qquad \text{and} \qquad p_2 = \frac{1}{2}\, e_X\,(e_X - 1) - C_X \quad (2.23)$$

$\Pr[Z_s Z_t = 1 | \mathcal{C}_1]$ is the probability that any pair of randomly chosen edges shares a node, and this is just the ratio of the number of edges with a common node in $\mathcal{G}_Y$, to the number of edge pairs in the complete graph with N nodes.

As $\mathcal{G}_Y = \cup_{j=1}^{K} \mathcal{K}_{n_j}$, the degree of each node on this graph is $n_j - 1$, and for $\mathcal{K}_N$, it is $N - 1$.

The latter graph has $e_{\mathcal{K}_N} = \binom{N}{2}$ edges. For these two graphs, the numbers of pairs of edges sharing a node are:

$$C_Y = \frac{1}{2}\sum_{j=1}^{K} n_j (n_j - 1)(n_j - 2)$$
$$C_{\mathcal{K}_N} = \frac{1}{2} N(N-1)(N-2) \tag{2.24}$$

So, if edges s and t have a common node:

$$\Pr[Z_s Z_t = 1 \mid C_1] = \frac{C_Y}{C_{\mathcal{K}_N}} = \frac{2\,C_Y}{N(N-1)(N-2)}. \tag{2.25}$$

If edges s and t do not share a node, we have

$$\Pr[Z_s Z_t = 1 \mid C_2] = \frac{\sum_{j=1}^{K} \varepsilon(n_j) \sum_{k=1}^{K} \varepsilon(n_k - 2\,\delta_{jk})}{e_{\mathcal{K}_N}\, e_{\mathcal{K}_{N-2}}} \tag{2.26}$$

where $\varepsilon(n)$ denotes the number of edges in a complete graph with n nodes, and δ_{jk} is Kronecker's delta.

Equation (2.26) is simply the ratio of the number of disjoint edges in $\mathcal{G}_Y$ to the number of disjoint edges in the complete graph $\mathcal{K}_N$. After some algebra, it can be written as:

$$\Pr[Z_s Z_t = 1 \mid C_2] = \frac{4\,e_Y(e_Y - 1) - 8\,C_Y}{N(N-1)(N-2)(N-3)}. \tag{2.27}$$

Hence, combining these results with equation (2.22), we can write the variance of Γ_R as:

$$\text{var}\,(\Gamma_R \mid e_X, e_Y, C_X, C_Y) =$$

$$\mu_1 + 2\left\{\frac{2\,C_X\,C_Y}{N(N-1)(N-2)} + (P_X - C_X)\frac{4\,e_Y(e_Y-1) - 8\,C_Y}{N(N-1)(N-2)(N-3)}\right\} - \mu_1^2$$

$$= \frac{2\,e_X\,e_Y}{N(N-1)}\left\{1 - \frac{2\,e_X\,e_Y}{N(N-1)}\right\} + \frac{4}{N(N-1)(N-2)}$$

$$\cdot\left[C_X\,C_Y + \frac{\{e_X(e_X-1) - 2\,C_X\}\,\{\{e_Y(e_Y-1) - 2\,C_Y\}}{N-3}\right] \qquad (2.28)$$

In the particular case where $\mathcal{G}_Y = \mathcal{K}_N$, we have that $e_Y = \binom{N}{2}$, and $C_Y = \frac{1}{2}N(N-1)(N-2)$, and, consequently, $\mu_1 = e_Y$, and $\text{var}\,(\Gamma_R) = 0$. This is what we should expect, as there would be no variation in the degree sequence of $\mathcal{G}_Y$ once that N and the sample sizes are fixed.

Steele et al. (1987) proved that for the 1-*MST* constructed with the usual Euclidean distance the number of nodes of any degree tends to a constant which depends only on the dimension of the space where the nodes lie. This result has an application in the context of multivariate runs tests. We now enunciate the main theorem of the paper by Steele et al.

Theorem 5 *If X_i, $1 \leq i \leq \infty$ are i.i.d. with density f in $\mathbb{R}^p$, and if $V_{k,N}$ denotes the number of nodes of degree k in an 1-MST with N nodes, then, with probability 1:*

$$\lim_{N\to\infty} V_{k,N} = \alpha_{k,p}$$

for $p \geq 2$ and $k \geq 1$

As half the sum of the degrees on 1-*MST* equals to $N-1$, we have that:

$$C_X = \frac{1}{2} \sum_{i=1}^{N} \deg_i (\deg_i - 1) = \frac{1}{2} \sum_k k^2 V_{k,N} - N + 1$$

As μ_2 depends on $\mathcal{G}_X$ only through e_X and C_X, and this last parameter is asymptotically independent of the topology of $\mathcal{G}_X$, then the variance of Γ_R, and thus, also the tests based on a Normal approximation to the distribution of Γ_R, are, w.p. 1, asymptotically unconditional on the topology of $\mathcal{G}_X$, when this graph is the 1-*MST*.

Important as it is, Steele et al.'s result has little practical application in the context of the multivariate tests we are interested in, because the constants $\alpha_{k,p}$ are unknown, except for a few particular cases, and its proof depends crucially on some geometrical properties of the 1-*MST* which do not seem liable to be extended for other kinds of graphs.

The third central moment is obtained using the expressions:

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1 + 2\mu_1^3 \tag{2.29}$$

with μ_2' and μ_1 as in expressions (2.21) and (2.19), and

$$\mu_3' = \mathrm{E}\left(\sum_{i=1}^{e_X} Z_i\right)^3$$

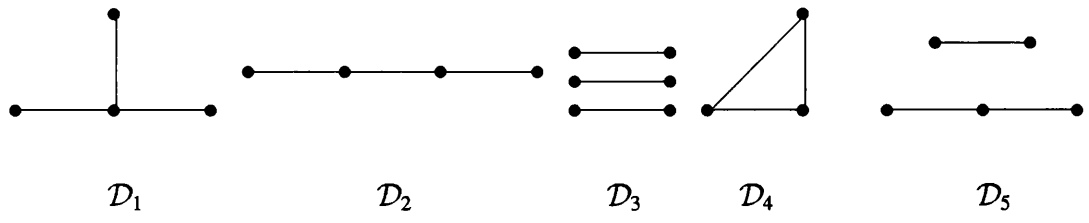
$$= \mathrm{E}\left(\sum_{i=1}^{e_X} Z_i^3 + 3\sum_{i\neq j}^{e_X} Z_i^2 Z_j + 6\sum_{i<j<k}^{e_X} Z_i Z_j Z_k\right)$$

$$= \mu_1 + 3\sum_{i<j}^{e_X} \mathrm{E}(Z_i Z_j) + 6\sum_{i<j<k}^{e_X} \mathrm{E}(Z_i Z_j Z_k). \tag{2.30}$$

We can now calculate $\sum_{i<j}^{e_X} \mathrm{E}(Z_i Z_j)$ from equations (2.25) and (2.27).

The value of $\sum_{i<j<k}^{e_X} \mathrm{E}(Z_i Z_j Z_k)$ depends on the form in which any three edges are linked (or not) on a graph. Now we calculate this expectation conditioning on the five

configurations that can be formed with three different edges, which are (Cliff and Ord, 1981):



$\mathcal{D}_1$ $\mathcal{D}_2$ $\mathcal{D}_3$ $\mathcal{D}_4$ $\mathcal{D}_5$,

so

$$\sum_{i<j<k}^{e_X} \mathrm{E}\,[Z_i Z_j Z_k] = \sum_{m=1}^{5} q_m \Pr[Z_s Z_t Z_u = 1 \mid \mathcal{D}_m] \qquad (2.31)$$

where q_m is the observed number of groups of three edges appearing as in configuration $\mathcal{D}_m$ within $\mathcal{G}_X$ and s, t and u are any three edges in $\mathcal{G}_Y$. $\Pr[Z_s Z_t Z_u = 1 \mid \mathcal{D}_m]$, is the ratio of the number of configurations $\mathcal{D}_m$ that one can form in $\mathcal{G}_Y$ to the corresponding number of such configurations found in the complete graph $\mathcal{K}_N$.

So we have

$$\Pr[Z_s Z_t Z_u = 1 \mid \mathcal{D}_1] = \frac{\sum_{j=1}^{K} n_j (n_j - 1)(n_j - 2)(n_j - 3)}{N(N-1)(N-2)(N-3)}$$

$$\Pr[Z_s Z_t Z_u = 1 \mid \mathcal{D}_2] = \frac{\sum_{j=1}^{K} \varepsilon(n_j)(n_j - 2)(n_j - 3)}{e_{\mathcal{K}_N}(N-2)(N-3)}$$

$$\Pr[Z_s Z_t Z_u = 1 \mid \mathcal{D}_3] = \frac{\sum_{j=1}^{K} \varepsilon(n_j) \left\{ \sum_{k=1}^{K} \varepsilon(n_k - 2\delta_{jk}) \left[\sum_{l=1}^{K} \varepsilon(n_l - 2\delta_{jl} - 2\delta_{kl}) \right] \right\}}{e_{\mathcal{K}_N} e_{\mathcal{K}_{N-2}} e_{\mathcal{K}_{N-4}}}$$

$$\Pr[\, Z_s Z_t Z_u \; = 1 \,|\, \mathcal{D}_4 \,] \;=\; \frac{C_Y}{C_{\mathcal{K}_N}}$$

$$\Pr[\, Z_s Z_t Z_u \; = 1 \,|\, \mathcal{D}_5 \,] \;=\; \frac{\sum_{j=1}^{K} C_{\mathcal{K}_{n_j}} \sum_{k=1}^{K} \varepsilon(n_k - 3\,\delta_{jk})}{C_{\mathcal{K}_N}\, e_{\mathcal{K}_{N-3}}}$$

It is not possible to obtain simple expressions for the q_ms in terms of the degree sequence of $\mathcal{G}_X$ as it was done for p_1, p_2 in equation (2.23), and so this numbers have to be calculated by direct enumeration over the edges of $\mathcal{G}_X$.

Combining the expressions obtained for $\Pr[\, Z_s Z_t Z_u \; = 1 \,|\, \mathcal{D}_m \,]$ for the five possible configurations $\mathcal{D}_m$ with equations (2.30) and (2.31), we can now calculate μ_3:

$$\mu_3(\,\Gamma_R \,|\, e_X, e_Y, \{p_m\}, \{q_m\}\,) = \mu_1 + 6\mu_2' - 3\mu_2'\mu_1 + 2\mu_1^3 + 6 \sum_{m=1}^{5} q_m \Pr[\, Z_s Z_t Z_u \; = 1 \,|\, \mathcal{D}_m \,]$$

The fourth central moment is obtained using the following identities:

$$\mu_4 \;=\; \mu_4' \;-\; 4\mu_3'\mu_1 \;+\; 6\mu_2'\mu_1^2 \;-\; 3\mu_1^4 \qquad (2.32)$$

and

$$\mu_4' \;=\; \mathrm{E}\left(\sum_{i=1}^{e_X} Z_i \right)^4$$

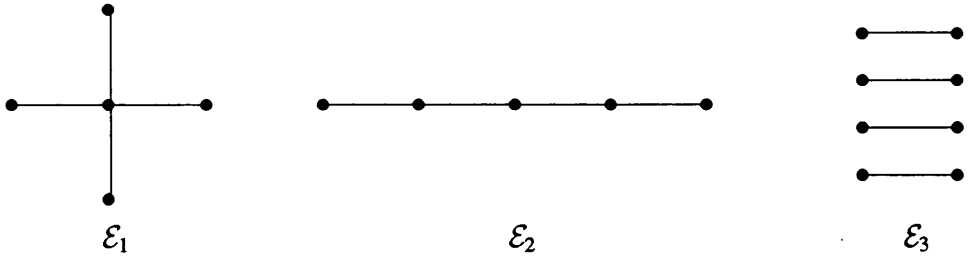
$$= \;\mathrm{E}\left(\sum_{i=1}^{e_X} Z_i^4 \;+\; 4 \sum_{i \neq j}^{e_X} Z_i^3 Z_j \;+\; 6 \sum_{i<j}^{e_X} Z_i^2 Z_j^2 \;+\; 12 \sum_{\substack{i \neq j,k \\ j<k}}^{e_X} Z_i^2 Z_j Z_k \right.$$

$$\left. +\; 24 \sum_{i<j<k<l}^{e_X} Z_i Z_j Z_k Z_l \right)$$

$$= \mu_1 + 14 \sum_{i<j}^{ex} E(Z_i Z_j) + 36 \sum_{i<j<k}^{ex} E(Z_i Z_j Z_k)$$

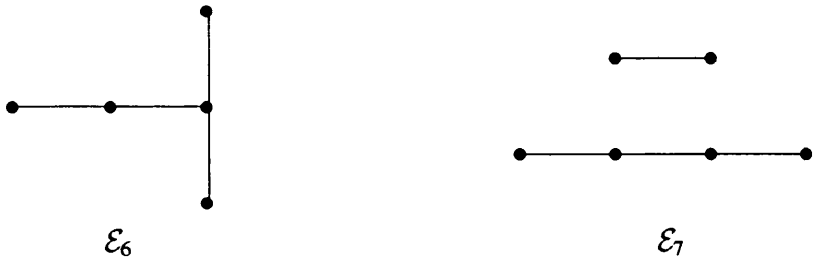
$$+ 24 \sum_{i<j<k<l}^{ex} E(Z_i Z_j Z_k Z_l) \qquad (2.33)$$

Except for the last expected value, all the terms in the last formula can be calculated using equations (2.30), (2.21) and (2.19). In order to obtain $\sum_{i<j<k<l}^{ex} E(Z_i Z_j Z_k Z_l)$, we have to consider 11 forms of combining 4 different edges within a graph (Cliff and Ord, 1981); these are:

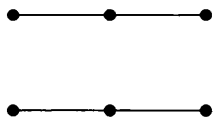


$\mathcal{E}_1$ $\mathcal{E}_2$ $\mathcal{E}_3$

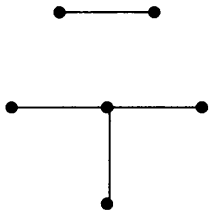
$\mathcal{E}_4$ $\mathcal{E}_5$



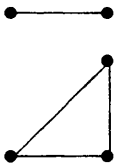
$\mathcal{E}_6$ $\mathcal{E}_7$



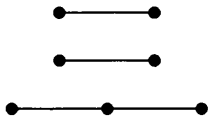
$\mathcal{E}_8$



$\mathcal{E}_9$



$\mathcal{E}_{10}$



$\mathcal{E}_{11}$

Next, we need to calculate

$$\sum_{i<j<k<l}^{e_X} \mathrm{E}\,[\,Z_i Z_j Z_k Z_l\,] \;=\; \sum_{m=1}^{11} r_m \,\Pr[\,Z_s Z_t Z_u Z_v \;=\; 1 \,|\, \mathcal{E}_m\,] \qquad (2.34)$$

for any four edges s, t, u and v in $\mathcal{G}_X$. The coefficients r_m are the observed numbers of four edges arranged as configurations $\mathcal{E}_m$ appearing within $\mathcal{G}_X$. We now proceeed using the same notation as used for the third moment.

$$Pr[\,Z_s Z_t Z_u Z_v \;=\; 1 \,|\, \mathcal{E}_1\,] \;=\; \frac{\sum_{j=1}^{K} n_j\,(n_j-1)\,(n_j-2)\,(n_j-3)\,(n_j-4)}{N\,(N-1)\,(N-2)\,(N-3)\,(N-4)}$$

$$\Pr[\,Z_s Z_t Z_u Z_v \;=\; 1 \,|\, \mathcal{E}_2\,] \;=\; \frac{\sum_{j=1}^{K} C_{\mathcal{K}_{n_j}}\,(n_j-3)\,(n_j-4)}{C_{\mathcal{K}_N}\,(N-3)\,(N-4)}$$

$$\Pr[\, Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_3 \,] =$$

$$\frac{\sum_{j=1}^{K} \varepsilon(n_j) \left\{ \sum_{k=1}^{K} \varepsilon(n_k - 2\delta_{jk}) \left[\sum_{l=1}^{K} \varepsilon(n_l - 2\delta_{jl} - 2\delta_{kl}) \left(\sum_{m=1}^{K} \varepsilon(n_m - 2\delta_{jm} - 2\delta_{km} - 2\delta_{lm}) \right) \right] \right\}}{e_{\mathcal{K}_N} e_{\mathcal{K}_{N-2}} e_{\mathcal{K}_{N-4}} e_{\mathcal{K}_{N-6}}}$$

$$\Pr[\, Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_4 \,] = \frac{\sum_{j=1}^{K} \varepsilon(n_j)\,(n_j - 2)\,(n_j - 3)}{e_{\mathcal{K}_N}\,(N-2)\,(N-3)}$$

$$\Pr[\, Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_5 \,] = \frac{\sum_{j=1}^{K} C_{\mathcal{K}_{n_j}}\,(n_j - 3)}{C_{\mathcal{K}_N}\,(N-3)}$$

$$\Pr[\, Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_6 \,] = \frac{\sum_{j=1}^{K} n_j\,(n_j-1)\,(n_j-2)\,(n_j-3)\,(n_j-4)}{N\,(N-1)\,(N-2)\,(N-3)\,(N-4)}$$

$$\Pr[\, Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_7 \,] = \frac{\sum_{j=1}^{K} \varepsilon(n_j) \sum_{k=1}^{K} \varepsilon(n_k - 2\delta_{jk})\,(n_k - 2 - 2\delta_{jk})\,(n_k - 3 - 2\delta_{jk})}{e_{\mathcal{K}_N}\, e_{\mathcal{K}_{N-2}}\,(N-4)\,(N-5)}$$

$$\Pr[\, Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_8 \,] = \frac{\sum_{j=1}^{K} C_{\mathcal{K}_{n_j}} \sum_{k=1}^{K} C_{\mathcal{K}_{n_k - 3\delta_{jk}}}}{C_{\mathcal{K}_N}\, C_{\mathcal{K}_{N-3}}}$$

$$\Pr[\,Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_9\,] = \frac{\sum_{j=1}^{K} \varepsilon(n_j) \sum_{k=1}^{K} \varepsilon(n_k - 2\delta_{jk})\,(n_k - 2 - 2\delta_{jk})\,(n_k - 3 - 2\delta_{jk})}{e_{\mathcal{K}_N}\, e_{\mathcal{K}_{N-2}}\,(N-4)\,(N-5)}$$

$$\Pr[\,Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_{10}\,] = \frac{\sum_{j=1}^{K} \varepsilon(n_j) \sum_{k=1}^{K} C_{\mathcal{K}_{n_k - 2\delta_{jk}}}}{e_{\mathcal{K}_N}\, C_{\mathcal{K}_{N-2}}}$$

$$\Pr[\,Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_{11}\,] = \frac{\sum_{j=1}^{K} C_{\mathcal{K}_{n_j}} \sum_{k=1}^{K} \varepsilon(n_k - 3\delta_{jk}) \sum_{l=1}^{K} \varepsilon(n_l - 5\delta_{jk} - 2\delta_{kl})}{C_{\mathcal{K}_N}\, e_{\mathcal{K}_{N-3}}\, e_{\mathcal{K}_{N-5}}}$$

So the fourth central moment can be written as:

$$\begin{aligned}\mu_4(\Gamma_R \mid e_X, e_Y, \{p_m\}, \{q_m\}\,\{r_m\}) &= \mu_1 + 14\mu_2' + 36\mu_3' - 4\mu_3'\mu_1 + 6\mu_2'\mu_1^2 - 3\mu_1^4 \\ &\quad + 24 \sum_{m}^{11} q_m \Pr[\,Z_s Z_t Z_u Z_v = 1 \mid \mathcal{E}_m\,]. \qquad (2.35)\end{aligned}$$

The skewness and kurtosis measures are defined as:

$$\begin{aligned}\beta_1 &= \frac{\mu_3^2}{\mu_2^3} \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} \\ \gamma_1 &= \sqrt{\beta_1} \qquad \gamma_2 = \beta_2 - 3\end{aligned} \qquad (2.36)$$

Unfortunately, the computation of the coefficients $\{q_m\}$ and $\{r_m\}$, which are needed to obtain the third and fourth central moments can be very expensive. There is no way of calculating these coefficients as a simple function of the degree sequence of $\mathcal{G}_X$, as it was done for the configurations involving only one or two edges. The only possibility is to enumerate the different configurations encountered within $\mathcal{G}_X$. This task can be easily achieved for sparse graphs with relatively few nodes, or graphs, as the first *MST* and the first *NNG*, with a maximum degree D_p^* depending only on p and

bounded independently of N, such that $D_p^* \ll N$. The number of operations involved in the enumeration procedure is at most proportional to the fourth power of the maximum degree of $\mathcal{G}_X$. As this quantity can be, in many cases, of almost the same order of magnitude as $N/2$, the computational burden of these calculations may be too heavy. Another consequence of this fact is that it makes impossible to produce general expressions that allow us to evaluate how fast do β_1 and β_2 converge to 0 and 3, respectively. However, the values of these measures for any particular case may still be regarded as useful summaries of how close the corresponding null distribution is to its limiting distribution.

Semyaticki (1978) studied a similar problem for the Mantel-Valand space-time clustering statistics. For these clustering procedures (as well as for *MRPP*-type statististics) it is necessary to perform $r N^3$ operations in order to calculate the moment of order r. Semyaticki presents a technique to break down the expressions involved in the straightforward calculation of the higher moments into a series of patterns, each of which can be calculated relatively quickly. However, the number of operations needed in order to calculate the first four moments is still $\mathcal{O}(N^3)$. In a similar line of work, Mielke, Berry and Wong (1986) presented an algorithm to obtain the first four moments of *MRPP*-type statistics using Semyaticki's technique.

In order to construct approximations to the null distribution of Γ_R, we followed two approaches, discussed in the next section. One involves generating a sample of random permutations of the sample identities over the nodes of $\mathcal{G}_X$ in order to calculate the value of Γ_R for each permutation: this procedure approximates the null distribution, as under the null hypothesis all the permutations of sample identities have the same probability. The other uses the values of the third and fourth moments obtained in this section. The latter method gives a better insight of the null distribution of Γ_R; however, its computational cost may be too expensive for graphs which have nodes with relatively high degrees.

2.6 Approximations to the Null Distribution of Γ_R

In this section we describe two approximations to the null distribution of Γ_R. These procedures should be used whenever the sample sizes are large enough to make the computation of the complete exact permutational distribution virtually impossible and small enough to cast doubts about the adequacy of using the asymptotic null distribution. The first method simply computes the values of Γ_R for a large number of permutations of sample labels and obtains the significance level of the test by ranking the observed value of the statistic amongst its values calculated for the permutations. Another approach consists of fitting Pearson distributions based on the first three or four moments calculated under the null hypothesis. We now describe both procedures.

2.6.1 Sampling from the Exact Permutational Distribution of Γ_R

It is always possible to obtain the exact permutational distribution of a *GCC*-based statistic. For Γ_R, this is calculated with the permutations of the sample labels over the observed spanning graph $\mathcal{G}_X$. If there are K samples, n_j denotes the sample size for the j-th sample, and $N = \sum_{j=1}^{K} n_j$, then the number of permutations of these values is given by the multinomial coefficient

$$M = \binom{N}{n_1\, n_2\, \ldots\, n_K} = \frac{N!}{\prod_{j=1}^{K} n_j!}$$

and so, unless the total sample sizes are rather small, it is not feasible to enumerate all the possible permutations of the sample labels. An example showing two of all the possible sample labels permutations over an *MST* appears in Figure 2-1.

For every assignment of the sample identities we obtain a value of Γ_R. The significance level α is equal to the proportion of these values which are less or equal to the observed value of Γ_R.

Berry (1982) presented an algorithm that generates all the possible permutations of N objects considered $n_1, n_2, \ldots, n_K$ at a time. It is based on the following result:

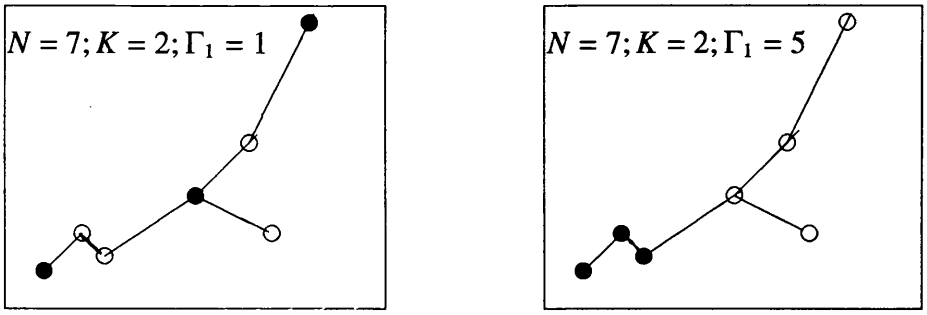


Figure 2-1: Two sample label permutations over $\mathcal{G}_X$

$$\binom{N}{n_1\, n_2\, \ldots\, n_K} = \binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{N-\sum_{j=1}^{K-2} n_j}{n_{K-1}}$$

The ordered list of computed values of Γ_R, and thus, the exact distribution of Γ_R for all permutations over $\mathcal{G}_Y$, depends only on the available data, avoiding any other assumptions. Some applications of Berry's algorithm can be seen in the papers by Mielke et al. (1982) and Zimmerman et al. (1985).

Figure 2-2 shows two examples of complete exact permutational distributions of Γ_R. The graphs used here were the first and the third *MST*s calculated on two bivariate standard normal samples, each of size 6; the multinomial coefficient is 924. The skewness and kurtosis measures β_1 and β_2 are 0.002 and 2.771 for the first distribution and 1.027 and 4.588 for the second one. It is possible to appreciate that even for such small sample sizes, the normal distribution seems to be an acceptable approximation for the first *MST*.

Unfortunately, unless the sample sizes are very small, it is not feasible to compute the value of the statistic for all the possible sample labels permutations. For example, if $N = 15, K = 3$, and $n_j = 5 (j = 1, 2, 3)$, we have $M = 756756$; for $N = 20, K = 5$ and $n_j = 4 (j = 1, \ldots, 5)$, M is 3.055×10^{11}, that is, about 400000 times the value for the previous sample sizes.

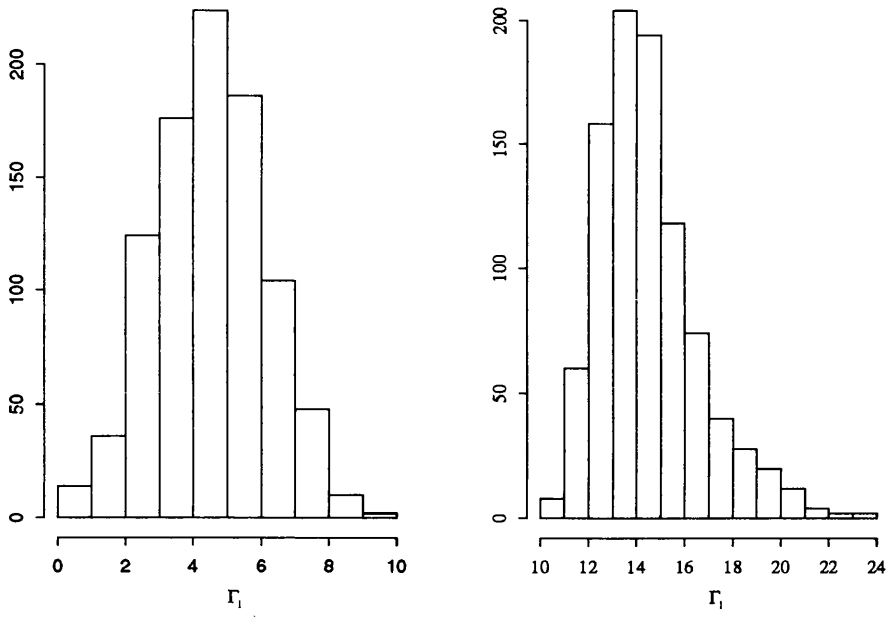


Figure 2-2: Exact Permutational Distributions; 1-MST and 3-MST

In such cases, we have to generate permutations of the integers

$$\{\underbrace{1,\ldots,1}_{n_1},\underbrace{2,\ldots,2}_{n_2},\ldots,\underbrace{K,\ldots,K}_{n_K}\}.$$

To generate a sample of permutations of the sample identities over $\mathcal{G}_X$, we used the following algorithm, due to Moses and Oakford (1963):

For $t = N$ down to 1 **do**

1. Generate s uniform in $\{1,\ldots, t\}$;
2. Swap the sample identities of the s-th and t-th objects.

Some properties of this algorithm have been studied by Ripley (1987).

In the following discussion we shall assume that the null hypothesis is rejected for relatively large values of the test statistic. If the null distribution of interest is continuous, then an unbiased estimator, $\hat{p_\alpha}$ of the true significance level is calculated with the proportion of values of the statistic produced by the sample of permutations which are less than the observed value of the statistic. The variance of this estimator is controlled by the number of permutations considered.

If the null distribution is discrete, so that ties may occur, a conservative procedure is given by computing the nominal significance level using the proportion of sampled values which are less than the observed value of the statistic. This ranking produces an upper bound to the real Significance level and has been used by Diggle (1983).

We adopt a policy that attempts to make some adjustment for the ties that occur due to the discrete nature of Γ_R. Hope (1968) suggested the following correction. Let p_1 be the p-value produced by the ranking described in the previous paragraph. If there are a number, say m, of the M sampled permutations which produced the same value of Γ_R as the one observed, then the corrected p value would be

$$p_\alpha = p_1 + \frac{m-1}{2M}$$

2.6.2 Pearson Distributions

The coefficients β_1 and β_2 are needed to fit a Pearson type curve to the distribution of Γ_R. A complete study of these curves is the subject of the book by Elderton and Johnson (1969). An abridged discussion, followed by some examples appears in the texts by Johnson and Kotz (1970) and by Kendall, et al. (1987, v.I). As these authors recall, the fitting of Pearson distributions is based on the estimation method of moments. This can lead to serious drawbacks if the observed data are a random sample from a population, as the method of moments does not yield, in general, efficient estimators of the population parameters. However, if the purpose is to obtain an expression which approximates a sampling distribution whose first four moments are known, as in this case, the method is usually satisfactory.

The authors mentioned in the previous paragraph discuss other possibilities to construct approximations to theoretical sampling distributions. For instance, one could represent a density function as a series in the derivatives of the normal density function, as in the Gram-Charlier and Edgeworth expansions. Another alternative is to seek for a transformation of the distribution of the variate into a known form (e.g. Johnson distributions). Although both approaches may be more flexible than using Pearson's curves given the first four moments, they certainly are more cumbersome to obtain and

use. In addition, we found three recent references (Berry, Mielke and Wang (1986), Tracy and Tajuddin (1986), and Tracy and Khan (1987)) in which Pearson curves were fitted to a nonparametric multivariate test statistic (*MRPP*, as defined by Mielke *et al.* (1976)) similar to Γ_R with excellent results. We decided to follow this approach to approximate the sampling distribution of Γ_R. A brief description of the Pearson distributions used follows.

Any pdf f which belongs to the Pearsonian system of distributions satisfies a differential equation of the form

$$\frac{df}{dx} = \frac{f(x)(x-a)}{b_0 + b_1 x + b_2 x^2} \tag{2.37}$$

The shape of f depends on the parameters a, b_0, b_1, and b_2. The above equation suggests that as its derivative vanishes at some point ($x = a$), f has a single mode, although there are particular solutions to equation (2.37) that lead to J–shaped or U–shaped distributions.

On the other hand, we also see that df/dx tends to 0 when f does so. The denominator of the r.h.s. of equation (2.37) is the second order MacLaurin's expansion of the corresponding distribution function. Considering this expansion, and assuming, without loss of generality, that $\mu_1 = 0$, it is possible to write the following system:

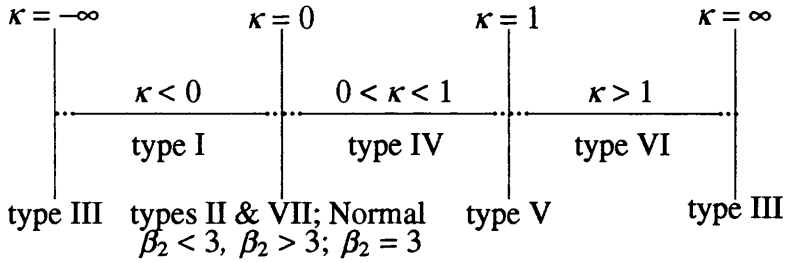
$$\begin{aligned}
a + b_1 &= 0 \\
b_0 + 3 b_2 \mu_2 &= -\mu_2 \\
a\mu_2 + 3 b_1 \mu_2 + 4 b_2 \mu_3 &= -\mu_3 \\
a\mu_3 + 3 b_0 \mu_2 + 4 b_1 \mu_3 + 5 b_2 \mu_4 &= -\mu_4
\end{aligned}$$

Using the solution to the previous equations it is possible to construct an expression in

terms of β_1 and β_2 that reflects the type of distribution which corresponds to the first four moments. This quantity, known as the *criterion*, and denoted by κ can be written as

$$\kappa = \frac{\beta_1 (\beta_2 + 3)^2}{4 (2\beta_2 - 3\beta_1 - 6) (4\beta_2 - 3\beta_1)} \tag{2.38}$$

If the roots of the quadratic $b_0 + b_1 x + b_2 x^2 = 0$ are both real with different signs, then $\kappa < 0$, and the type I curve is obtained. For complex roots, we have $0 < \kappa < 1$, corresponding to the type IV. Finally, for real roots of the same sign, $\kappa > 1$, and we get the type VI curve. These density functions are called the main Pearsonian types. Karl Pearson distinguished 12 types: some of them are trivial, some are no longer of interest. The other 9 functions, called transition types, correspond to the limiting situations when one of the main types changes into another. When $|\kappa|$ is large (theoretically, ∞), one root is ∞ (type III). If $\kappa = 1$, then both roots are equal (type V), and when $\kappa = 0$, the roots are equal in magnitude but with opposite signs (type II). For the case $\kappa = 0$ and $b_1 = b_2 = 0$, we obtain a curve which depends only on the first two moments: it is the Normal distribution; if $b_2 = 0$, Finally, if $\kappa = b_1 = 0$, the curve is Pearson's type VII, also known as Student's t distribution. The following diagram illustrates these points.



Another possibility for choosing the type of Pearson's curve needed is to locate the values of γ_1 and β_2 in the Johnson chart shown in Figure 2-3.

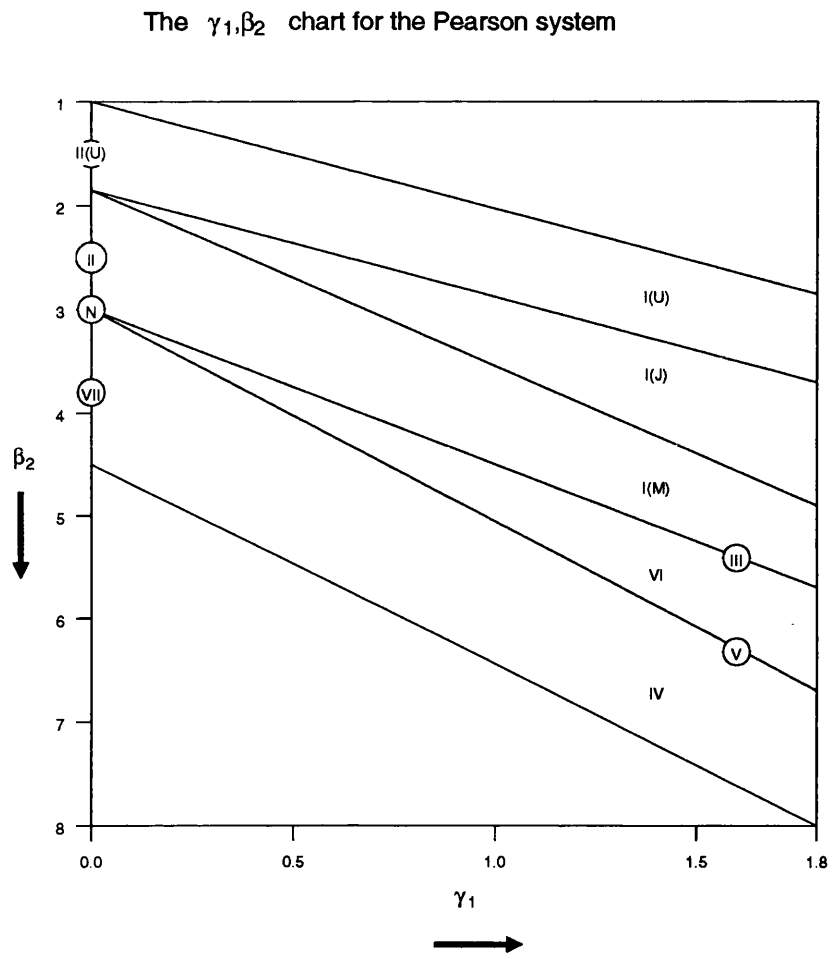


Figure 2-3: Johnson γ_1, β_2 chart

If g is the observed value of Γ_R, f is the chosen standarized Pearson density, and $g_0 = (g - \mu)/\sigma)$, then the approximate p-value is given by

$$\alpha = \int_{-\infty}^{g_0} f(x)\, dx$$

Davis and Stephens (1983) presented an algorithm to approximate the quantiles for eleven significance levels of a Pearson curve. For any given value of the statistic, the significance level can be easily approximated by interpolation. We used this algorithm to check the significance levels calculated via numerical integration, obtaining very good agreements. Berry et al. (1986), did not consider the transition curves types II and VII. In the cases in which these functions may be suited, they considered a normal distribution. However, we found some examples in which the types II and VII produced

slightly better fits to the exact permutational distribution than those obtained with the normal distribution, and accordingly, decided to work with them. We also found it useful to calculate significance levels based on type IV curves, thus leaving aside only the type V distribution.

The criterion κ should be used with caution to select a Pearson distribution, as it can vary quite drastically for values of β_1 and β_2 which are very near 0 and 3. We fitted a type IV curve only when $0.1 \leq \kappa \leq 0.9$. The choice of these bounds for κ was determined after having difficulties with the numerical integration routines almost always when κ was outside this interval; this situation is briefly mentioned by Elderton and Johnson (1969).

Figure 2-4 shows several examples of Pearson curves. The Normal density appears in the dotted lines in all the graphs.

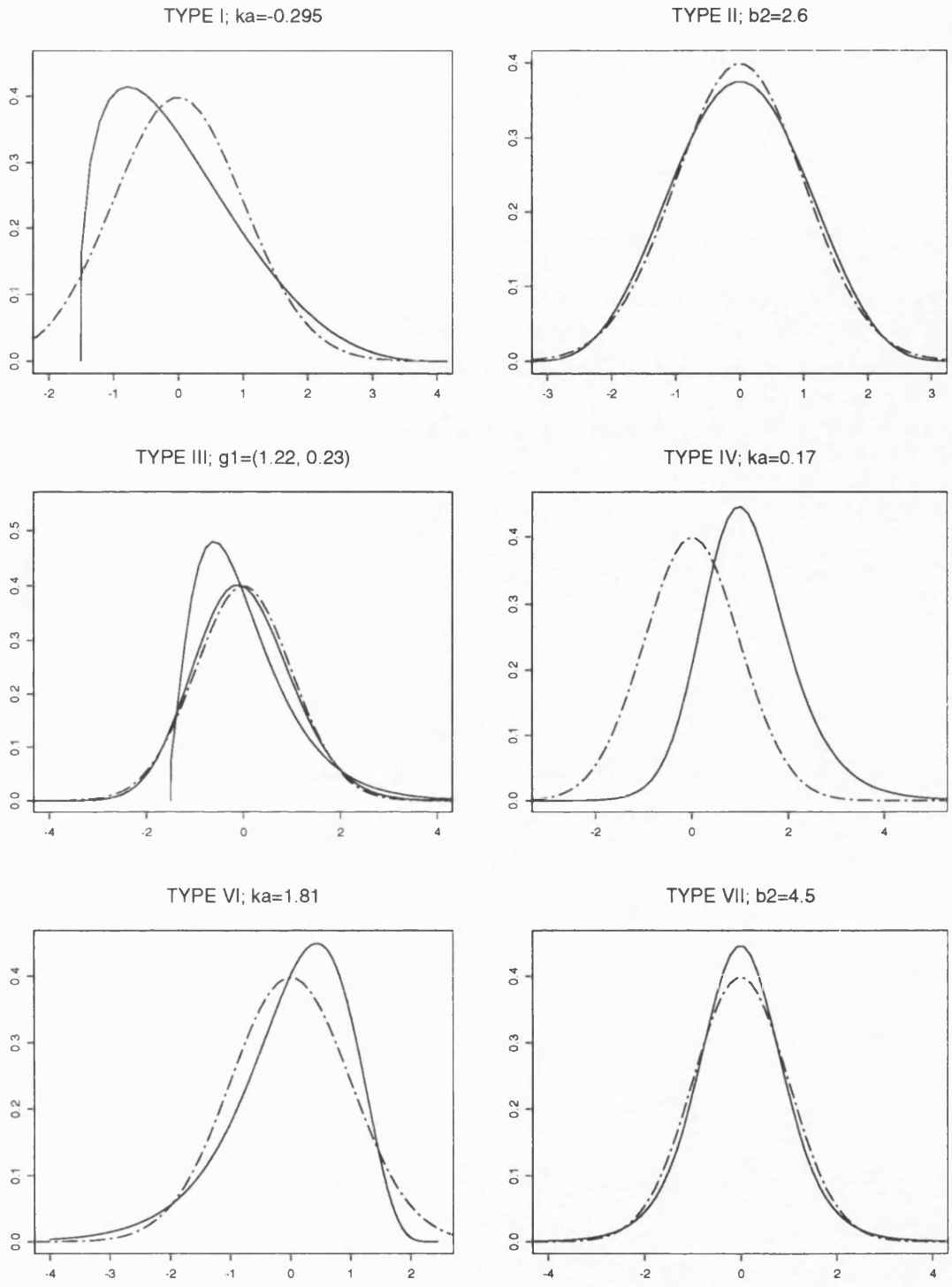


Figure 2-4: Pearson curves

2.7 Examples of the Approximations

In this section we discuss some examples of the approximations constructed using the methods of the last section. As it was noted, the calculation of Pearson curves can be computationally infeasible for certain graphs, as the number of operations to enumerate the number of configurations involving three and four edges is approximately proportional to the fourth power of the maximum degree found in $\mathcal{G}_X$. This is particularly critical while considering *GG*s for points in very high dimensional spaces or higher order *RNG*s or *GG*s. However, it is not a great problem for *n-NNG*s or *n-MST*s if the value of n is not very large. In other words, these approximations are a sensible option for runs tests based on a wide variety of graphs even for relatively large sample sizes. Thus, we aim to explore how well the approximations described in Section 2.6 perform for small sample sizes and to study how much may be gained in accuracy by using approximations based in the first four moments or in large samples from the permutational distribution instead of the Normal approximation for moderate sample sizes.

Problems with the asymptotic normality for the null distribution have been reported to appear in some *MRPP*-type statistics (Mielke et al. 1976). The choice of the weights used for defining the test statistic seems to be a crucial factor for the speed of convergence. Indeed, Mielke (1979) constructed a non-degenerate example in which the asymptotic distribution for an *MRPP* statistic is non-normal. Mantel and Valand (1970) and Semyaticki, (1978) also pointed out that the space-time clustering statistic proposed by Mantel (1967) needs very large sample sizes in order to regard the normal approximation as satisfactory. Constanzo et al. (1983) mentioned similar problems with respect to spatial autocorrelation statistics. The results of Subsection 2.5.2 allow us to assume the asymptotic normality of the null distribution of Γ_R. Of course, the choice of $\mathcal{G}_X$ and the sample structure do affect the validity of this approximation for any given total sample size.

In this section, we analyze these effects. In all the examples presented, we used the Euclidean distance calculated always on samples of p standarized independent

variables to construct $\mathcal{G}_X$. In several cases, we constructed the complete permutational distribution and calculated several descriptive statistics for samples of permutations of the sample identity over $\mathcal{G}_X$ with sizes between 10000 and 100000. An unbiased estimator, $\hat{p_\alpha}$, of the true significance level, is obtained by substracting the proportion of values of the statistic produced by the sample of permutations which are less than the observed value of the statistic from unity. The variance of $\hat{p_\alpha}$ decreases proportionally to the number of permutations considered (Scholz and Stephens, 1987).

Our first example considers relatively large sample sizes. We used the same sample sizes as Friedman and Rafsky (1979). We worked with two samples, each of 100 standard multivariate normal deviates in several dimensions, In Friedman and Rafsky's example, their runs statistic was calculated with the first three orthogonal *MST*s. It is virtually impossible to compute the whole exact permutational distribution for these examples, as the multinomial coefficient is of order 10^{29}, but, given the moderately large sample sizes, we might expect the normal approximation to perform reasonably. We generated two samples of size 100 from a bivariate standard normal distribution and calculated the value of Γ_R for 50000 random permutations of the sample identities over the first three *MST*s based on Euclidean distances. We did so using the method described in Subsection 2.6.1. The results for the 1-*MST* and the 3-*MST* appear in Figures 2-5 and 2-6, respectively.

The values of the skewness and kurtosis coefficients β_1 and β_2, calculated with the exact method of Subsection 2.5.3 conditional on the observed *MST*s, were (0.0003, 2.979) for 1-*MST* and (0.0067, 3.031) for 3-*MST*, which closely resemble those corresponding to a normal distribution.

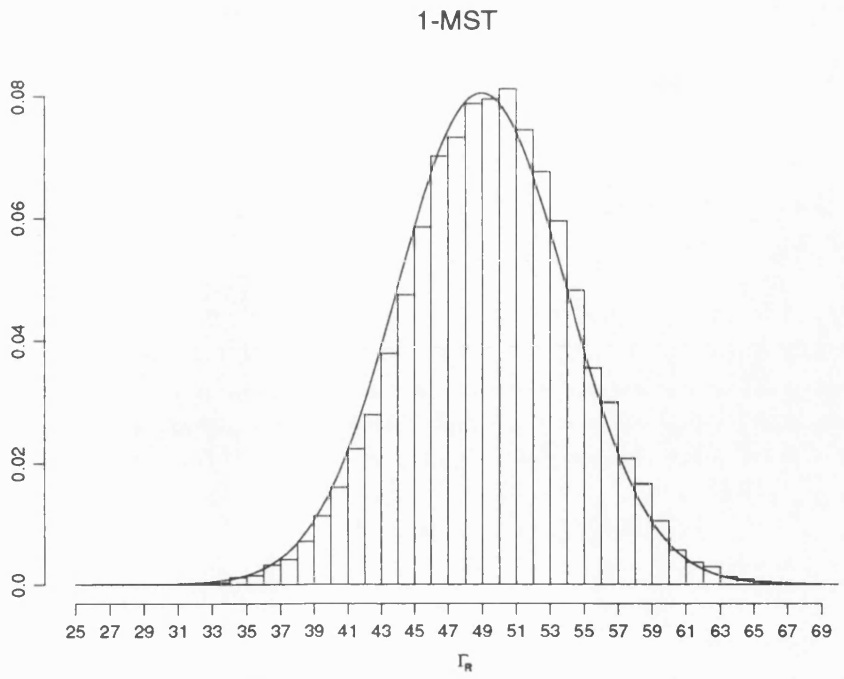


Figure 2-5: 1-*MST*: Permutational and Normal approximations

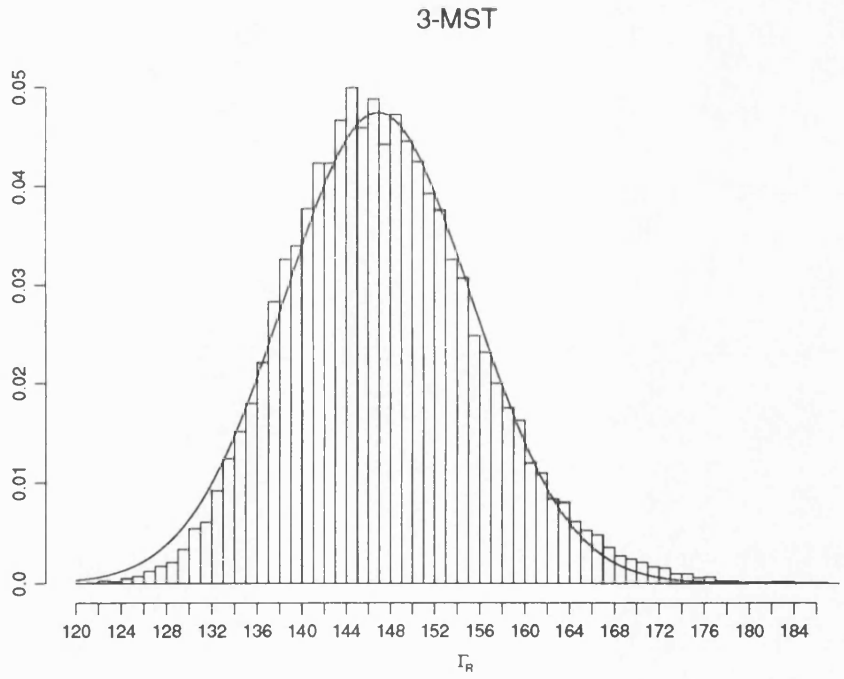


Figure 2-6: 3-*MST*: Permutational and Normal approximations

A further confirmation of our claim of normality for this distribution results from the significance levels of the goodness of fit statistics Z_1, Z_2 and K^2 proposed by D'Agostino et al. (1990).

Thus, Friedman and Rafsky were right in using the normal approximation for their example. However, this is not always the case, and a word of caution should be said here. For instance, Figure 2-7 shows the distribution of Γ_R over 50000 random permutations of the sample identities on the 6-*MST* for the same bivariate data. The Type IV approximation appears as the dotted line; the Normal approximation is the continuous line. As it can be seen, the percentage levels calculated with the Pearsonian curve (indicated as *P* in the figure) correspond much more closely to those of the sampled permutational distribution than the ones obtained with the Normal approximation. The difference is more substantial for higher quantiles.

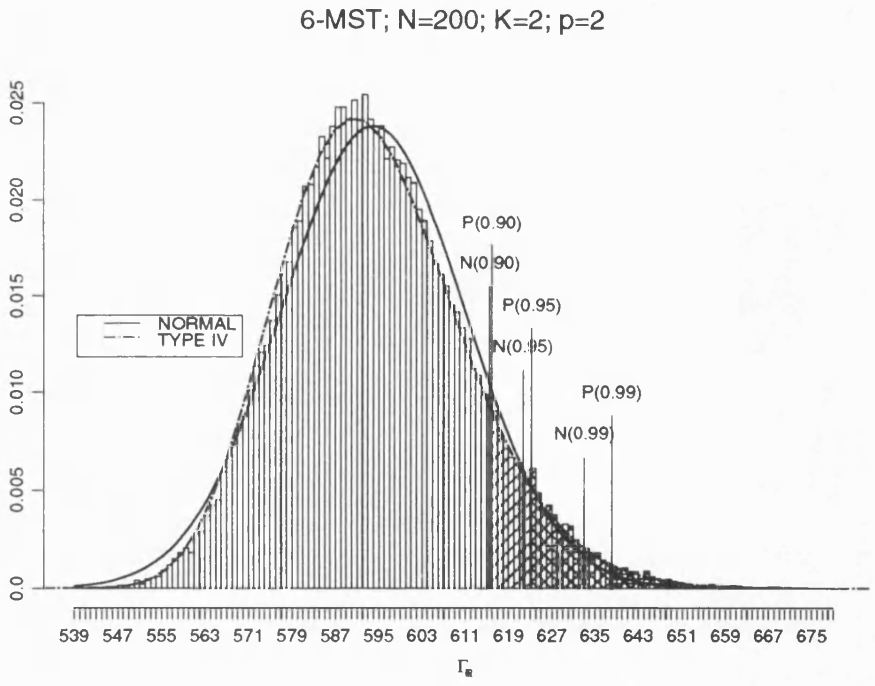


Figure 2-7: 6-*MST*: Permutational, Normal and Type IV approximations

For this distribution we have that $\beta_1 = 0.0221$ and $\beta_2 = 3.068$, calculated with the exact method. Although these values seem close enough to the corresponding values of the Normal distribution, the criterion κ equals 0.2396, clearly pointing towards a

Type IV approximation. This example shows that the Normal approximation should not be taken for granted even for total sample sizes as large as 200 for higher order orthogonal spanning graphs.

Table 2.1 shows the values of (β_1, β_2) corresponding to data in 5, 10, and 20 dimensions for several graphs calculated over two standard multivariate Normal samples, each of size 100.

The entries corresponding to the *GG*s which are marked with an asterisk were estimated from a sample of 100000 permutations of the sample identities, instead of using the exact method. This is so because the *GG* has a very high degree of connectedness, and, as a result, it is virtually impossible to enumerate the edge configurations required to calculate the exact third and fourth moments of the null distribution of Γ_R.

For our next example, we generated two samples, each of size 10, of bivariate independent normal random variables. For this sample structure there are 184756 possible permutations of the sample labels on the observed graph. We obtained several graphs based on Euclidean distances for this sample and calculated the first four moments; we also constructed the complete permutational distribution. The descriptive statistics calculated for this distribution are, of course, the same as those obtained using the method described in Subsection 2.5.3. The results appear in Table 2.2.

Table 2.1: (β_1, β_2) values; N=200; $n_1 = n_2 = 100$, under H_0

	$p = 2$			$p = 5$			$p = 10$			$p = 20$		
Graph	β_1	β_2	*links*	β_1	β_2	*links*	β_1	β_2	*links*	β_1	β_2	*links*
$1-MST$	0.00001	2.95938	199	−0.00002	2.99588	199	0.00000	2.98313	199	0.00003	3.02676	199
$2-MST$	0.02219	3.03135	298	0.00953	2.99728	298	0.00136	3.00957	298	0.00369	3.02803	298
$3-MST$	0.06315	3.15054	397	0.02906	3.00284	397	0.00676	3.03185	397	0.00647	3.03030	397
$4-MST$	0.09411	3.19569	496	0.04241	3.04352	496	0.01360	3.04135	496	0.00853	3.04112	496
$5-MST$	0.12996	3.25116	595	0.06549	3.09287	595	0.01836	3.05428	595	0.01183	3.04493	595
$6-MST$	0.16565	3.31791	694	0.08882	3.17209	694	0.02225	3.06839	694	0.01604	3.05147	694
$1-NNG$	0.00000	2.98551	138	0.00000	2.98549	138	0.00000	2.98655	148	0.00000	2.98741	158
$2-NNG$	0.00804	3.00026	255	0.00301	2.99707	272	0.00193	2.99748	292	0.00015	2.99604	300
$3-NNG$	0.02254	3.02891	375	0.00679	3.01041	399	0.00449	3.00678	4243	0.00082	2.99916	432
$4-NNG$	0.04156	3.06495	495	0.01358	3.02724	522	0.00720	3.01702	555	0.00234	3.00352	567
$5-NNG$	0.05673	3.09609	600	0.02395	3.05038	647	0.01139	3.02740	691	0.00425	3.00986	706
$6-NNG$	0.07559	3.13331	715	0.03477	3.07470	772	0.01567	3.03867	824	0.00595	3.01560	840
$1-RNG$	0.00000	2.99272	241	−0.00001	2.99748	320	−0.00006	2.99858	437	−0.00016	2.99896	530
$2-RNG$	0.04022	3.06797	612	0.05782	3.11555	1189	0.05654	3.11472	2123	0.05757	3.11181	3189
$1-GG$	0.00484	3.00911	364	0.05107	3.16583	1436			5175			14720
$2-GG$	0.10083	3.19837	1055	0.27382*	3.78745*	8295			19584			19900

Parameter values					
Graph	μ	σ	β_1	β_2	κ
$1-MST$	9.000	2.126	0.000	2.893	0.000
$2-MST$	18.000	2.790	0.158	3.330	0.667
$3-MST$	27.000	3.157	0.449	3.998	0.578
$4-MST$	36.000	3.254	0.568	4.162	0.787
$5-MST$	45.000	3.244	0.412	4.073	0.376
$6-MST$	54.000	3.108	0.479	4.075	0.564
$1-NNG$	7.579	1.987	0.001	2.884	–0.002
$2-NNG$	13.263	2.527	0.241	3.238	–0.776
$3-NNG$	20.368	2.977	0.344	3.527	11.588
$4-NNG$	24.632	3.183	0.459	3.776	2.198
$5-NNG$	31.263	3.372	0.683	4.193	1.784
$6-NNG$	36.474	3.412	0.828	4.541	1.256
$1-RNG$	9.000	2.126	0.000	2.893	0.000
$2-RNG$	20.842	2.880	0.459	3.683	–32.042
$3-RNG$	48.316	3.141	1.104	5.093	1.214
$4-RNG$	84.789	1.476	0.266	4.760	0.081
$1-GG$	16.105	2.708	0.044	3.111	0.365
$2-GG$	41.684	3.378	0.728	4.612	0.623
$3-GG$	79.105	2.028	0.340	4.036	0.264

Table 2.2: N=20, $n_1 = n_2 = 10, p = 2$, under H_0

For the first order graphs, the values of β_1 and β_2 are close to 0 and 3, respectively, and thus, indicate that normal approximation would be adequate, even for such small sample sizes. This does not happen for higher order graphs, where an approximation based in four moments should be used. A different picture emerges from keeping everything but the number of dimensions constant. Table 2.3 shows the results obtained by considering 10 dimensions instead of 2.

It does seem that the normal approximations work better for higher dimensional data for almost all the graphs considered.

This pattern was consistently observed for several configurations with unequal sample sizes for $N = 10, 20$ and number of variables between 2 and 20. The *RNG* and the *GG* for orders higher than 3 and 1, respectively, were the complete graph, with 190 links. In this case, the variance of Γ_R equals 0, and thus, it is not possible to calculate the exact β_1 and β_2.

Parameter values					
Graph	μ	σ	β_1	β_2	κ
$1-MST$	9.000	2.059	0.001	2.869	–0.003
$2-MST$	18.000	2.633	0.017	2.939	–0.074
$3-MST$	27.000	3.048	0.032	2.965	–0.147
$4-MST$	36.000	3.254	0.059	2.949	–0.161
$5-MST$	45.000	3.301	0.069	3.137	0.796
$6-MST$	54.000	3.187	0.104	3.035	–0.330
$1-NNG$	7.105	1.861	0.001	2.891	–0.005
$2-NNG$	14.211	2.497	0.009	2.947	–0.164
$3-NNG$	20.368	2.888	0.030	2.976	–0.164
$4-NNG$	26.526	3.138	0.056	2.945	–0.153
$5-NNG$	30.316	3.290	0.088	3.015	–0.287
$6-NNG$	37.895	3.319	0.062	2.971	–0.194
$1-RNG$	13.263	2.382	–0.012	2.978	1.293
$2-RNG$	45.000	3.063	0.412	3.789	0.998
$3-RNG$	87.158	1.039	0.000	2.826	0.001
$1-GG$	72.947	2.635	0.000	3.015	–0.005

Table 2.3: N=20, $n_1 = n_2 = 10$, $p = 10$, under H_0

Tables 2.4 and 2.5 show the results obtained for 50 multivariate standard normal variates, divided in 5 samples, with the following sizes $n_1 = 8, n_2 = 8, n_3 = 15, n_4 = 10, n_5 = 9$, and considering $p = 2$ and $p = 20$. In this case, the multinomial coefficient is 1.08×10^{31}, so we obtained 100000 permutations of the sample labels over the observed graph and calculated the value of Γ_R in each case. These tables show the values of the mean, standard deviation, β_1 and β_2 calculated as in Subsection 2.5.3 and those that we got from the sampled permutational distribution.

As it can be seen, the agreement is, in general, very good. This was also the case for similar data in dimensions 5 and 10.

As a further illustration of some of the approximations, we now present the results obtained for one of the series of 10000 permutations from the exact distribution for 6-*NNG* in 2 and 20 dimensions. Again, we have $N = 50$ points sampled from a standard multivariate normal distribution, classified in 5 samples, with sizes $n_1 = 8, n_2 = 8, n_3 = 15, n_4 = 10, n_5 = 9$.

	Parameter values							
	$1-MST$		$1-NNG$		$1-RNG$		$1-GG$	
	theor.	perm.	theor.	perm.	theor.	perm.	theor.	perm.
μ	9.680	9.700	6.914	6.889	11.260	11.252	16.989	16.999
σ	2.726	2.740	2.317	2.350	2.936	2.934	3.555	3.559
β_1	0.042	0.048	0.067	0.085	0.035	0.033	0.055	0.040
β_2	3.002	3.104	3.010	3.055	3.001	3.027	3.087	3.028
	$6-MST$		$6-NNG$		$2-RNG$		$2-GG$	
μ	58.080	58.119	38.127	37.069	27.855	27.863	44.054	44.089
σ	5.951	6.011	5.071	5.057	4.450	4.420	5.380	5.371
β_1	0.203	0.186	0.152	0.188	0.094	0.089	0.125	0.106
β_2	3.422	3.361	3.327	3.451	3.195	3.152	3.268	3.283

Table 2.4: $N = 50; n_1 = 8, n_2 = 8, n_3 = 15, n_4 = 10, n_5 = 9; p = 2$, under H_0

	Parameter values							
	$1-MST$		$1-NNG$		$1-RNG$		$1-GG$	
	theor.	perm.	theor.	perm.	theor.	perm.	theor.	perm.
μ	9.680	9.651	7.902	7.915	20.743	20.764	222.245	222.234
σ	2.729	2.737	2.478	2.452	3.902	3.911	3.831	3.778
β_1	0.039	0.035	0.049	0.037	0.008	0.009	–0.011	–0.013
β_2	2.992	2.959	2.992	2.923	2.969	3.008	2.986	2.934
	$6-MST$		$6-NNG$		$2-RNG$		$2-GG$	
μ	58.080	58.049	41.486	40.856	92.256	92.288	242	242
σ	5.992	5.983	5.265	5.255	6.826	6.874	0	0
β_1	0.027	0.015	0.031	0.034	0.030	0.019	0	0
β_2	3.048	2.954	3.053	3.029	3.046	3.078	0	0

Table 2.5: $N = 50; n_1 = 8, n_2 = 8, n_3 = 15, n_4 = 10, n_5 = 9; p = 20$, under H_0

In Figure 2-8 we show the results for bivariate standard normal data. The normal and the type IV approximations, together with their (0.90, 0.95 0.99) quantiles appear in the figure.

The type IV approximation follows the sampled permutational distribution slightly better than the normal one, particularly in the tails. The theoretical values of β_1 and β_2 were 0.152 and 3.327, respectively, giving a value of 0.598 for κ, thus confirming the adequacy of a Pearson type IV fit for the null distribution of Γ_R in this case.

Figure 2-9 shows the analogous situation for 20-dimensional normal observations. Although in rigour a Type VI distribution should be used, as the value of κ was 1.802, the advantages of doing so are negligible as the corresponding standard type VI distribution for the observed values of β_1 and β_2, (0.031 and 3.053) is undistinguishable for all practical purposes from the standard normal distribution.

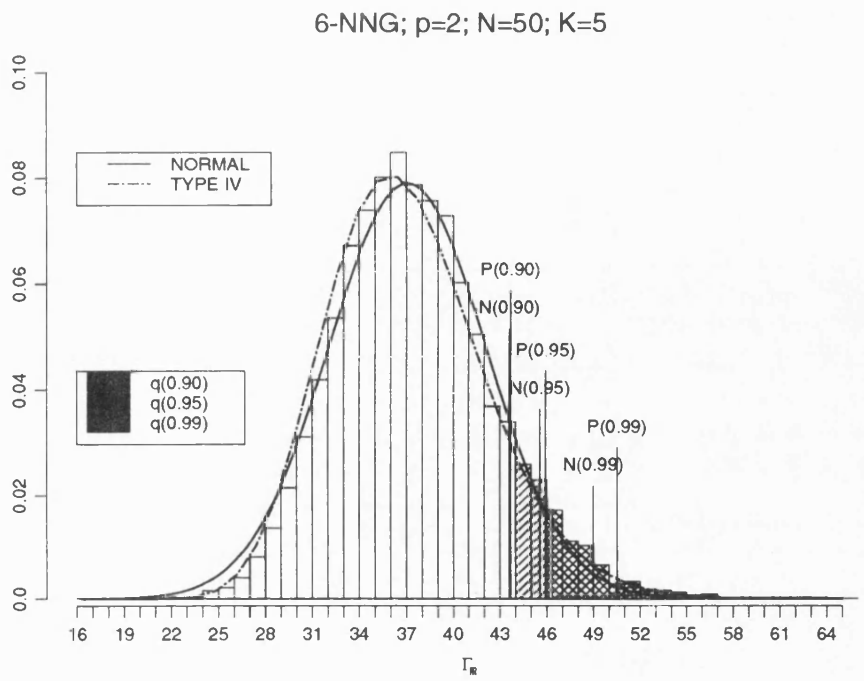


Figure 2-8: 100000 permutations and Normal and Type IV approximations

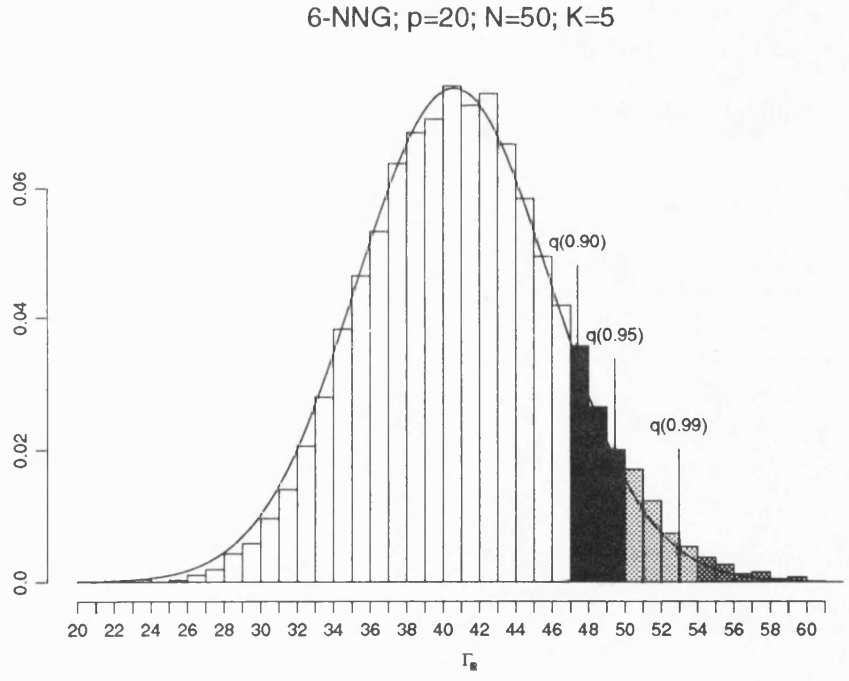


Figure 2-9: 100000 permutations and Normal and Type VI approximations

In our final example, we worked with three samples, each of size 9, of p-dimensional uniform distributions. Tables 2.6 and 2.7 show part of the results obtained for the exact four moments and the estimators calculated with the 100000 permutations of sample identities. The agreement between both approximations is, again, very good.

	Parameter values							
	$1-MST$		$1-NNG$		$1-RNG$		$1-GG$	
	theor.	perm.	theor.	perm.	theor.	perm.	theor.	perm.
μ	8.000	7.997	5.538	5.527	8.923	8.919	12.308	12.306
σ	2.330	2.333	1.971	1.968	2.449	2.452	2.813	2.814
β_1	0.023	0.022	0.042	0.041	0.019	0.017	0.028	0.029
β_2	2.937	2.942	2.929	2.937	2.953	2.961	3.016	3.020
	$6-MST$		$6-NNG$		$2-RNG$		$2-GG$	
μ	48.000	47.994	31.692	30.436	20.923	20.914	32.923	32.918
σ	4.225	4.215	4.011	3.929	3.488	3.488	3.893	3.895
β_1	0.594	0.597	0.425	0.447	0.193	0.197	0.183	0.185
β_2	4.144	4.179	3.808	3.886	3.339	3.318	3.354	3.359

Table 2.6: $N = 27; n_1 = 9, n_2 = 9, n_3 = 9; p = 2$, under H_0

	Parameter values							
	$1-MST$		$1-NNG$		$1-RNG$		$1-GG$	
	theor.	perm.	theor.	perm.	theor.	perm.	theor.	perm.
μ	8.000	8.001	6.462	6.462	12.615	12.609	100.00	100.001
σ	2.237	2.240	2.098	2.102	2.815	2.825	2.153	2.148
β_1	0.014	0.017	0.032	0.032	0.002	0.002	0.008	0.006
β_2	2.962	2.968	2.941	2.959	2.901	2.884	2.942	2.943
	$6-MST$		$6-NNG$		$2-RNG$		$2-GG$	
μ	48.000	47.996	32.000	31.454	43.692	43.682	0	0
σ	4.065	4.074	4.040	3.939	3.828	3.834	0	0
β_1	0.029	0.029	0.050	0.057	0.063	0.059	0	0
β_2	3.072	3.067	3.119	3.114	3.110	3.107	0	0

Table 2.7: $N = 27; n_1 = 9, n_2 = 9, n_3 = 9; p = 20$, under H_0

The examples in this section show that the normal approximation should be used carefully, and, if possible, avoided, except for lower order graphs and fairly large sample sizes. The approximations calculated fitting Pearson curves with the first four

moments or via a sample of random permutations are, in general, computationally feasible and produce good results.

The only pattern which is possible to extract from these tables is that the normal approximation has a better performance as the dimensionality of the data increases and that it works well for moderately small sample sizes.

Chapter 3

Multivariate Rank Tests

3.1 Introduction

We now turn to the second kind of tests of hypothesis outlined in the papers by Friedman and Rafsky. The basic idea is to obtain multivariate analogues of ranks in order to use them as input for univariate nonparametric procedures based on ranks. These ranks are the order in which the nodes in the 1-*MST* are visited when it is rooted at some particular node and then traversed. In the univariate case, to improve the power of the runs test against scale alternatives, one could rank the data with respect to the median of the pooled observations. For shift alternatives, one would use just the ordered list of observations. In the same way, for the multivariate ranking procedures the alternative hypothesis determines the root of the 1-*MST* as well as the traversing algorithm.

Friedman and Rafsky (1979) used two multivariate ranking procedures in conjunction with the univariate Smirnov two-sample test. They also suggested that their ranking methods could be used to generalize other nonparametric univariate rank tests. We investigate such generalizations and some of their applications in this section.

The context we are interested in is the construction of nonparametric tests for K samples of p-dimensional observations. Therefore we do not assume any parametric model for the distribution of the observations; we only assume that, under the null hypothesis, the K samples have the same distribution, say, F, and that $F \in \mathcal{F}$, where $\mathcal{F}$ is a class of distribution functions. The price we pay for this generality is a reduction of the

information used, achieved via multivariate rankings. An important point is to assess how well these ranking procedures reflect the relationships of closeness within the data distance matrix. Later in this chapter we address this problem and discuss some examples.

We begin this chapter with a thorough description of Friedman and Rafsky's ranking methods. We also give some examples in which we evaluate the performance of these methods for extracting the nearness relationships from the interpoint distances matrix. Next, we review the univariate versions of some nonparametric tests based on ranks that will be used together with the multivariate ranking procedures. These tests are the Smirnov (1939) two-sample test, the Kruskal-Wallis nonparametric *ANOVA*, a K-sample version of the normal scores test (Lehmann, 1975), one of Kiefer's (1959) K-sample tests, Conover's (1965) K-sample generalization of the Kolmogorov-Smirnov test and the Scholz-Stephens (1987) K-sample generalization of the Anderson-Darling goodness-of-fit test. We chose them among many univariate nonparametric multisample tests available from the literature because they have a distribution function, or an approximation to it, which is relatively easy to calculate. Finally, we discuss three multivariate nonparametric K-sample rank tests studied by Puri and Sen (1971). They are based on the ranks of the observations for each of the p-variables, thus differing from the approach suggested by Friedman and Rafsky.

3.2 Multivariate Analogues of Ranks

The main difficulty encountered for constructing multivariate generalizations of some well known univariate nonparametric tests lies in how to extend the concept of a sorted list of observations for multivariate data. To overcome this problem, it is possible to assign ranks to multivariate observations following the order in which the nodes of the 1-*MST* of the sampling points are visited in accordance to some traversing algorithm. For multivariate data, it is possible to construct nonparametric tests by conditioning on the observed interpoint distance graph. This is a situation analogous to that in the univariate case, when one conditions on the order statistics to obtain distribution free

test statistics.

We now describe two multivariate ranking methods proposed by Friedman and Rafsky (1979) and present some examples of their performance n inducing nearness relations on multivariate observations.

3.2.1 Diameter Ordering

The first ranking methods begins by constructing the 1-*MST* of the pooled sample and rooting it at a node with largest eccentricity, which is an end of a diameter of the 1-*MST*. The ranks of the points are then obtained as the order in which they are visited in a *height directed preorder traversal (HDPT)* of the 1-*MST*.

The *HDPT* algorithm for any tree can be defined recursively as:

1. visit the root;
2. HDPT in ascending order of height the subtrees rooted at the daughters of the root. (Resolve ties by visiting first subtrees with roots closer, in the distance measure used to construct the *MST*, to the node visited in step 1.)

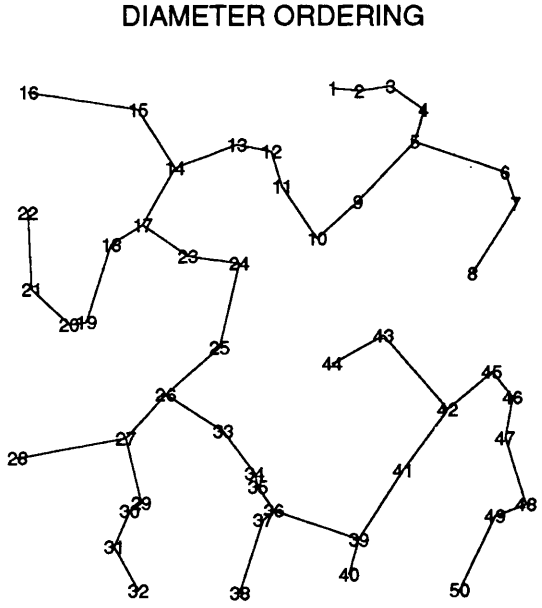


Figure 3-1: Multivariate diameter ranking based on 1-*MST*

An example of this ranking applied to 50 bivariate points appears in Figure 3-1. For univariate data, the 1-*MST* is equivalent to the sorted list. Therefore, this ranking may be seen as analogous to starting off the sorted list at one extreme of a univariate sample and ranking the observations according to their distances to that extreme. For K samples, location shifts in the multivariate space will be represented by concentrations within the ordered sequence of ranks of points of those samples which have a shift. Scale shifts for some samples should also appear in this sequence as scale shifts for the ranks corresponding to points from such samples.

3.2.2 Radial Ordering

The second ranking, called by Friedman and Rafsky *radial,* can be obtained as follows:

1. root the *MST* at the centre;
2. assign ranks such that nodes with larger depth receive higher ranks than those with smaller depth. Nodes with the same depth can be ordered in terms of their interpoint distance from the centre node.

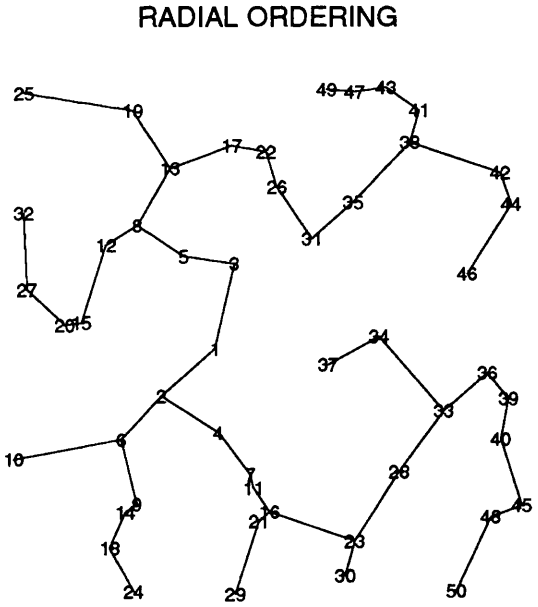


Figure 3-2: Multivariate radial ranking based on 1-*MST*

Figure 3-2 shows an example of this method for 50 points on the plane.

The node with largest eccentricity is not unique: there are at least two such nodes. Thus, there may be one or two centres of the *MST*. Friedman and Rafsky observed that the value of the Smirnov two sample statistic derived from this ranking does not change much by choosing any of the two possible centres.

The ranking produced with this method is analogous to ranking univariate observations with respect to their distances to the median of the observations, and so we would expect it to provide better power than the diameter ranking for alternatives based in differences in scale, rather than in location. Furthermore, as Zahn (1971) and Friedman and Rafsky (1981) pointed out, the edges of an *MST* tend to be directed along density gradients. Then, for spherically symmetric distributions (and their transformations), depth (rooting the tree at one centre) might be interpreted as the number of points on a steepest density descent path from the centre, and is an estimator of a quantity analogous to a weighted distance to the mode. Points located at portions of the radial ranks represent concentrations located at a given weighted distance from the centre, and so location differences appearing in this ranking correspond to scale differences in the multidimensional space.

3.2.3 Discussion

We now address the question of how well these multivariate ordering procedures reflect the nearness relationships which are present in the interpoint distance matrix. First we highlight some of the difficulties that such rankings have, considered as analogous to their univariate counterparts.

In the univariate case, the diameter ranking follows the data from one extreme of the pooled sample to the other. Analogously, we could think of assigning the lowest and highest ranks for points located in the extremes of the observations, while the ranks in the middle of the sequence $\{1, 2, \ldots, N\}$ would correspond to points closer to the centroid of the observations. For multivariate observations the two most extreme points in the periphery are not uniquely determined. Besides, the 1-*MST* diameter could be followed by the diameter ranking from one of the outermost regions of the data, to a central part; from there, to another peripheral region and then again to somewhere near

the centre of the data and so on, as illustrated in Figure 3-1.

The radial ranks for univariate data would begin with the median of the observations and, as it orders the observations according to their distances to the median, it may move towards any of the two extremes of the data. For multivariate observations, the lowest ranks should be assigned to points near the centre of the 1-*MST*, which is a plausible equivalent to the median of the data; the highest ranks should correspond to points in the periphery of the observations. This seems an easier task to achieve than the one specified for the diameter ranking, because it does not matter which one of the regions of the data are visited first, as long as we do so going from the inner towards the outmost parts, as it happens when we traverse the 1-*MST* radially. Figure 3-2 illustrates this point.

However, we cannot expect these analogies to perform in the same way as the univariate ordering. The two examples presented in Subsections 3.2.1 and 3.2.2 illustrate this. For higher dimensional data, the correspondence with the univariate rankings becomes weaker.

One way of evaluating how close these methods follow the rationale of univariate rankings involves the *convex hull* of the data. The convex hull of a set of observations is the minimal convex set that contains all the data. We can think of a sequence of convex hulls if we exclude the points forming the boundaries of previous convex hulls and construct the convex hull of the remaining observations. If we define the *convex hull depth* of each sample point by the number of convex hulls surrounding it, outermost points have depth 1, and interior points have relatively large depth values. Thus, if we plot the depth of each point against the ordered ranks, we should ideally expect to observe a sequence of curves going up (for points in the periphery of the data) and down (for points near the centroid of the observations) for the diameter ranking and a straight line with negative slope for the radial ranking.

There are several efficient algorithms to construct the convex hull for bivariate observations, e.g. Green and Silverman (1978), but, unfortunately, the convex hull is very complicated to obtain for higher dimensional points. As Bowyer (1981) and Watson (1981) noted, this is a dual problem to the p-dimensional Dirichlet tessellation

problem. Algorithms like the one presented by Jóźkis (1983) are suitable only for multivariate points satisfying certain geometrical conditions. Smith and Jain (1984) presented an heuristic procedure to decide if a point is or is not inside the convex hull generated by a given set of points. Although this technique may be adequate to approximate convex hulls, is not possible to use it in a straightforward manner to construct convex hulls.

We now present some examples of using the convex hull in order to assess the performance of the Friedman-Rafsky rankings for bivariate distributions.

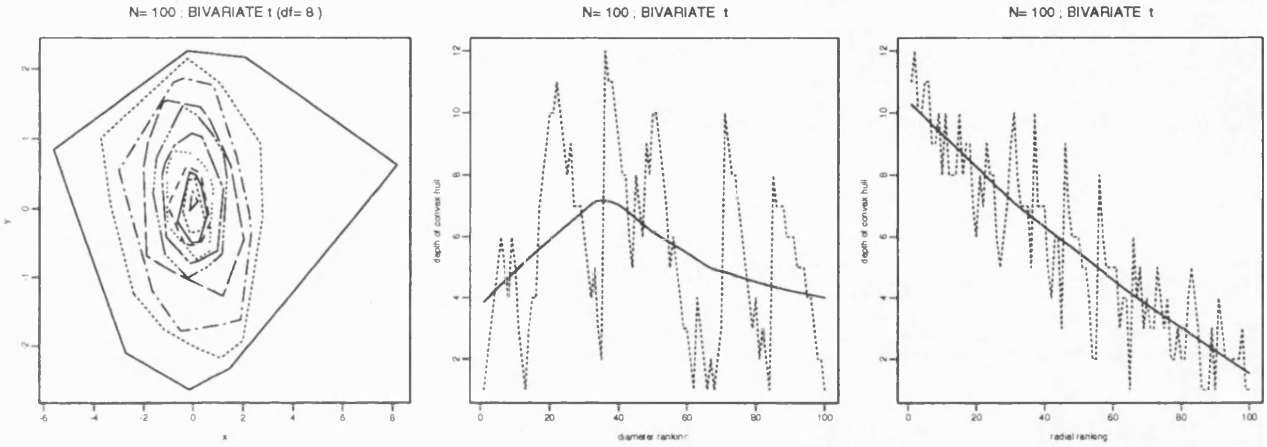


Figure 3-3: Convex hull depths: Bivariate $t(8)$

Figure 3-3 shows the sequence of convex hulls and the plots of convex hull depths against the two classes of ranks for two bivariate data drawn from two independent Student's t distributions with 8 degrees of freedom. A local scatterplot smoother (Cleveland, (1979)) has been added to the plots, as a continuous curve. It may be seen that the diameter ranking begins and ends with points located in the first layer of the convex hulls. In this case, the ranking behaves in accordance with the pattern mentioned before: it goes from outside points to inner points and from there to outer points, several times, with one large gap between two points in the first layer. The radial ranking follows the strata of convex hulls as we should expected.

However, this procedure of appraising the effectiveness of the rankings would not be

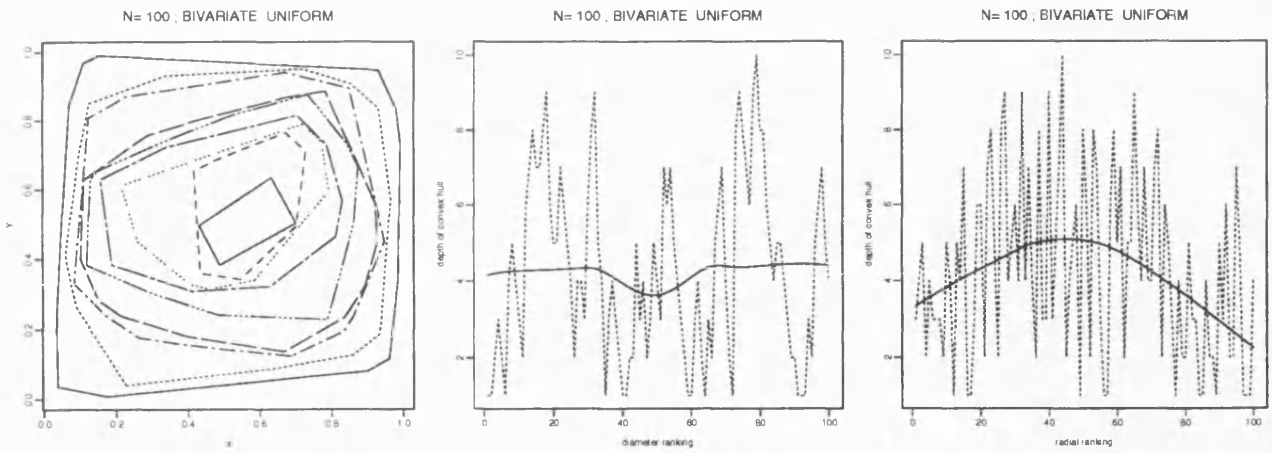


Figure 3-4: Convex hull depths: Bivariate uniform

very informative for distributions whose convex hulls layers are not sensitive enough to the position of a point in relation to the centroid of the observations. For example, Figure 3-4 shows the results for the bivariate uniform distribution.

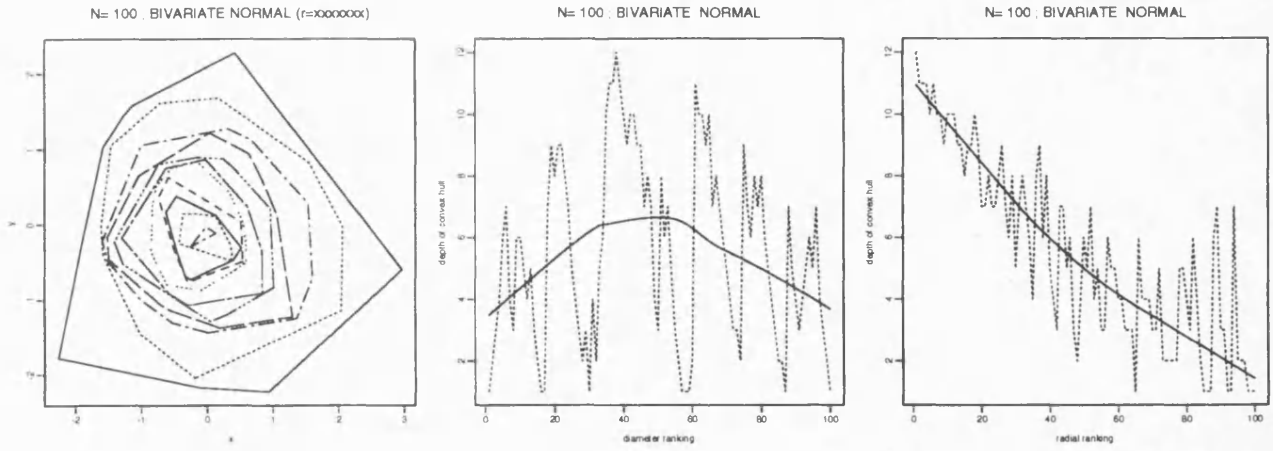


Figure 3-5: Convex hull depths: Bivariate normal

As another example, we show the corresponding plots for a bivariate normal distribution in Figure 3-5.

We should expect that as the dimension of the observations is increased the nearness

relations become more complicated, and thus the efficacy of the multivariate ranking methods decreases. Some possible criteria to evaluate to what extent the proposed rankings reflect the nearness structure of the data are:

- **Diameter ranking:**

    1. there should be a direct relationship between these ranks and the ranks of the distances of each point to one end of the 1-*MST*
    2. the majority of the distances between consecutive ranked points should be kept relatively small for all the ranks

- **Radial ranking:**

    1. there should be positive correlation between these ranks and the ranks of the distances of each point to the centre of the 1-*MST*
    2. the distances between consecutive ranked points should have an increasing trend.

We now present some examples. In them, the top two plots represent the ranks of the distances of each point to the end and to the centre of the 1-*MST* against the diameter and radial rankings, respectively. Scatterplot smoothers (continuous curves), as well as regression lines (dash-dotted lines) were added to these plots. The third plot is a density estimate of the places that the $(N-1)$ distances between consecutively diameter-ranked points occupy among all the ordered distances, divided by $\binom{N}{2}$. Finally, on the right-hand lower corner, we plotted, as a dotted curve, the position that each of the $(N-1)$ distances between consecutively radial-ranked points occupy, divided by $\binom{N}{2}$, against the ordered radial ranks; a scatterplot smoother also appears on the plot, as a continuous curve.

Figure 3-6 shows a case in which the criteria outlined above are satisfied. The correlation coefficients between both multivariate rankings and the ranks of the distances to the end and to the centre of the *MST* are highly significant. The first plot shows that the highest ranks are assigned to the points which are furthermost from the

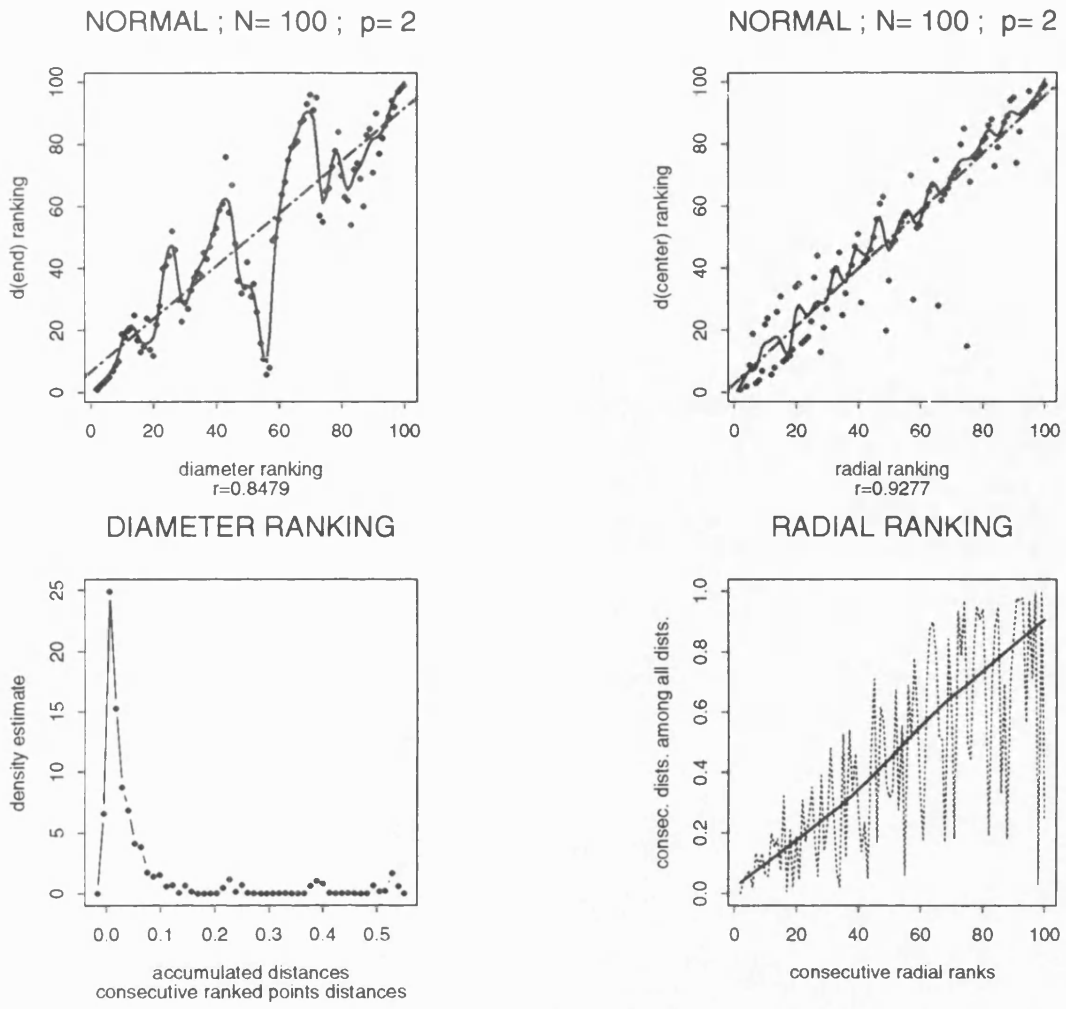


Figure 3-6: Multivariate rankings; Bivariate Normal

end, as it happens in the one-dimensional ranking situation. However, some of the points which are closer to the end of the diameter were assigned ranks between 50 and 60.

The correlation is even stronger for the radial ranking, with only a few points deviating from the general behaviour. Most (72%) of the distances between consecutive ranked points are among the 5% smallest elements of the distance matrix, and those distances do not increase for higher diameter ranks. Finally we see a very clear trend in the fourth plot, indicating that the radial ranks are assigned to points (or small clusters of points) in such a way that the point with the next rank is, in general, far away from the previous point.

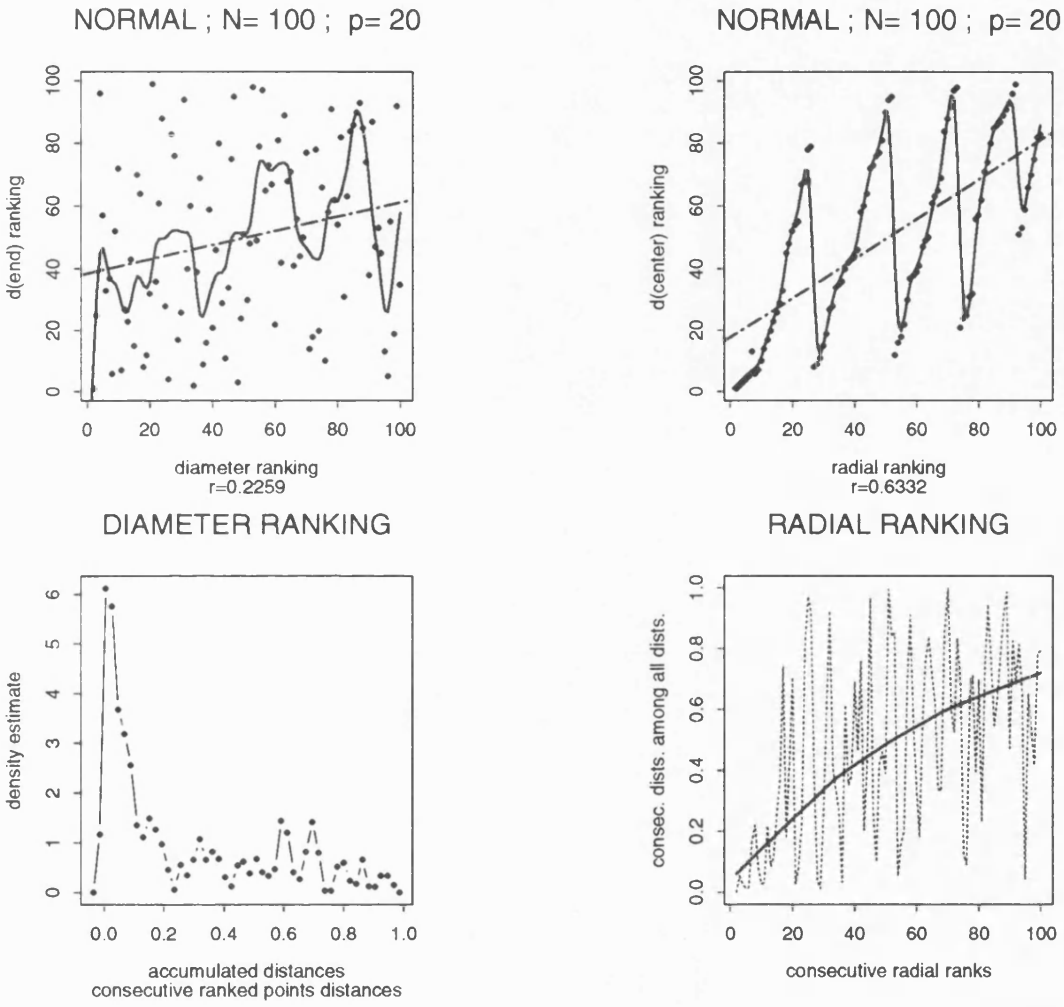


Figure 3-7: Multivariate rankings; Multivariate Normal, (p=20)

The situation changes drastically for a 20-dimensional normal distribution (Figure 3-7), as the 'curse of dimensionality' takes its toll. For the diameter ranking, we see that the association between the ranks and the distance to the end of the 1-*MST* follows a rather chaotic pattern, though the correlation is still non negligible. The radial ranks are associated with the distances to the centre in a stronger way. Five well defined branches of the *MST* are clearly marked in the plot, indicating the presence of directions which the data tend to follow.

The proportion of distances between consecutively diameter-ranked points which are among the 5% smallest distances was about a third, while for bivariate data such proportion was almost 90% of such small distances. However, still the vast majority of those distances are within the first 20% smallest distances. Again, a clear increasing trend is found in the relative positions of the distances between consecutive radial-ranked points in the sorted list of all the distances.

An example of these plots for a sample of size 200 of 30 independent Lognormal random variables appears in Figure 3-8. The same features observed for the 20-dimensional normal distribution appear in this example, perhaps even more clearly. Again we observe a high correlation between the diameter ranks and the distances to the end of the 1-*MST*, as well as between the radial ranks and the distances to the centre of the 1-*MST*. Almost a third of the distances between consecutive diameter-ranked points were among the 5% smallest distances. There is a clear direct relationship between the radial ranks and the distances between consecutive radial-ranked points.

Again, several branches of the 1-*MST* appear in the plot of the distances to the centre against the radial ranks; they are also marked in the consecutive radial ranks plot.

Another way of evaluating how well a multivariate diameter ranking mirrors its univariate analogue is looking at the distribution of the order of the distances that appear between consecutive ranked points. We carried out several simulations comparing the performance in this respect of the diameter ranking with the projections of the data points into the first principal component, which may be seen as another kind of multivariate ranking procedure, albeit a rather coarse one.

Table 3.1 gives an idea of the distribution of the distances between consecutively ranked

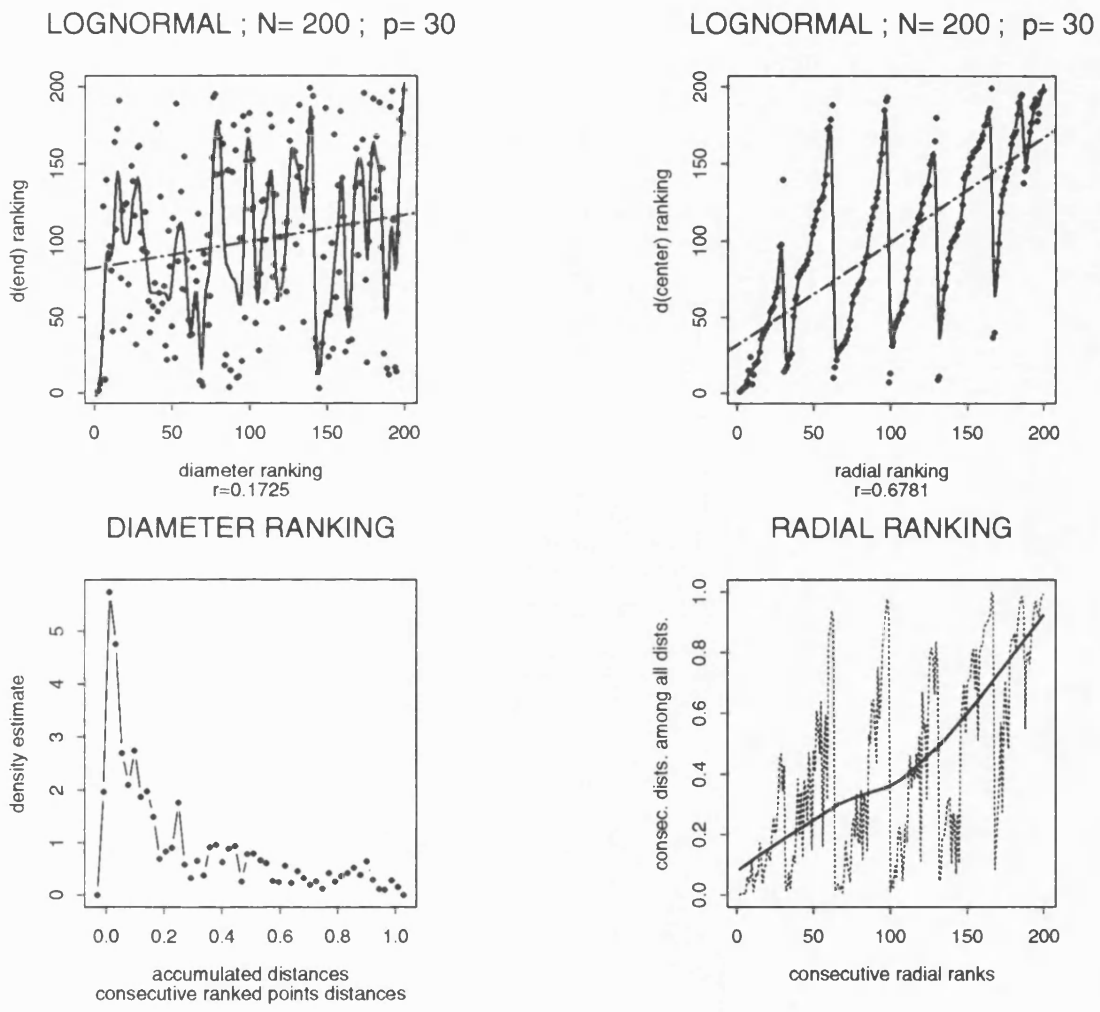


Figure 3-8: Multivariate rankings; Multivariate Lognormal, (p=30)

points.

In this case, we generated 100 samples of several combinations of dimensions, total sample size and data distribution. The results presented here are the averages, for the 100 samples, of the proportions of the $N-1$ distances obtained between consecutively ranked points which were among the smaller q percent of all the $\binom{N}{2}$ distances. The numbers in parentheses are the corresponding proportion of distances generated by ordering the observations projected into the first principal component.

These examples point out how, as the dimension of the data increases, both ranking procedures compared become more and more indadequate to produce an ordering of multivariate observations which preserves the nearness between contiguously ranked points. However, the diameter ranking consistently produces better results (in terms of the number of the smallest distances preserved for contiguous ranks) than the ranking obtained by projecting the multivariate observations into their first principal component. The figures presented in Table 3.1 are similar to those obtained for other sample sizes.

Both rankings induced using the 1-*MST* may be used as input for any univariate nonparametric rank test. Obviously, the power of the resulting tests would depend on the size of the discrepancy between the nearness relations induced by the resulting ranks and those present in the complete interpoint distance matrix. As Friedman and Rafsky (1979) pointed out, for tests against location alternatives, the diameter ordering would provide the best choice of ranks for higher power tests; the radial ordering being more suitable for scale alternatives. We discuss the power of the tests that result from applying these ranking methods to univariate nonparametric rank tests in Chapter 5.

In the next section we review the univariate tests used.

3.3 Univariate Rank Tests

3.3.1 Smirnov Test

The Smirnov (1939) test, also known as the two-sample Kolmogorov-Smirnov test, is based on the discrepancy between the two sample empirical distribution functions. Let

Distances between consecutive ranked points
diameter ranking and first principal components (N=100)

	Normal				Lognormal			
q	2	5	10	20	2	5	10	20
0.05	0.742	0.568	0.458	0.385	0.644	0.337	0.248	0.163
	(0.245)	(0.126)	(0.096)	(0.084)	(0.256)	(0.125)	(0.089)	(0.074)
0.10	0.849	0.698	0.588	0.505	0.771	0.433	0.338	0.246
	(0.360)	(0.212)	(0.174)	(0.164)	(0.373)	(0.213)	(0.164)	(0.138)
0.25	0.931	0.847	0.753	0.665	0.888	0.590	0.507	0.424
	(0.569)	(0.435)	(0.363)	(0.349)	(0.562)	(0.416)	(0.349)	(0.318)
0.50	0.971	0.946	0.897	0.827	0.950	0.748	0.699	0.640
	(0.776)	(0.691)	(0.639)	(0.612)	(0.769)	(0.662)	(0.606)	(0.587)
0.75	0.992	0.988	0.972	0.947	0.978	0.876	0.853	0.825
	(0.921)	(0.875)	(0.855)	(0.832)	(0.908)	(0.856)	(0.823)	(0.809)
0.90	0.999	0.998	0.995	0.986	0.988	0.947	0.939	0.927
	(0.979)	(0.963)	(0.949)	(0.946)	(0.971)	(0.952)	(0.943)	(0.932)
0.95	1.000	1.000	0.999	0.995	0.994	0.974	0.971	0.961
	(0.991)	(0.984)	(0.979)	(0.977)	(0.987)	(0.975)	(0.980)	(0.969)

	Exponential				Uniform			
q	2	5	10	20	2	5	10	20
0.05	0.697	0.538	0.422	0.309	0.849	0.715	0.605	0.523
	(0.269)	(0.133)	(0.101)	(0.081)	(0.288)	(0.149)	(0.113)	(0.098)
0.10	0.819	0.676	0.552	0.433	0.929	0.815	0.704	0.615
	(0.383)	(0.223)	(0.185)	(0.146)	(0.408)	(0.236)	(0.194)	(0.174)
0.25	0.922	0.825	0.730	0.627	0.975	0.916	0.926	0.738
	(0.589)	(0.423)	(0.371)	(0.334)	(0.621)	(0.448)	(0.396)	(0.375)
0.50	0.969	0.916	0.871	0.810	0.993	0.992	0.980	0.956
	(0.792)	(0.674)	(0.636)	(0.598)	(0.811)	(0.702)	(0.659)	(0.627)
0.75	0.988	0.962	0.943	0.928	0.999	0.998	0.996	0.986
	(0.924)	(0.871)	(0.843)	(0.826)	(0.929)	(0.881)	(0.858)	(0.839)
0.90	0.995	0.984	0.981	0.978	1.000	0.998	0.996	0.986
	(0.976)	(0.961)	(0.949)	(0.936)	(0.974)	(0.965)	(0.952)	(0.946)
0.95	0.997	0.993	0.991	0.991	1.000	1.000	0.999	0.994
	(0.991)	(0.986)	(0.977)	(0.972)	(0.992)	(0.983)	(0.978)	(0.977)

Table 3.1: Contiguously ranked distances

X_{ji} denote the i-th value of the j-th sample; then, the *edf* of the n_j observations in the j-th sample evaluated at any point x is simply the number of sample values of X_{ji} which are less than or to x divided by the corresponding sample size, n_j:

$$F_{X_j}(x) = \frac{1}{n_j} \cdot \#\{i \,|\, (X_{ji} \leq x,\ 1 \leq i \leq n_j)\} \tag{3.1}$$

The Smirnov two-sample test statistic is defined as follows:

$$D = \sup_x |\, F_{X_1}(x) - F_{X_2}(x) \,| \tag{3.2}$$

and H_0 is rejected for large values of D. Smirnov showed that under H_0, for large sample sizes n_1 and n_2, D follows the distribution that bears his name:

$$\Pr_{H_0}\left[D > z\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}\right] \approx 2\sum_{n=1}^{\infty}(-1)^{n-1}\, e^{-2n^2z^2} \tag{3.3}$$

The test is consistent for the same conditions as those mentioned for the runs test, and the statistic D follows the Smirnov distribution for any assignment of integer ranks of the N observations, as long as the ranking used does not involve the sample identities.

3.3.2 Kruskal-Wallis Test

The Wald-Wolfowitz univariate runs test (reviewed in Chapter 2) and the Smirnov test were designed as general alternatives tests. Another possibility is to construct tests for specific alternative hypotheses. The Kruskal-Wallis (1954) nonparametric analysis of variance is an adequate test if one is interested in contrasting the hypothesis of homogeneity against alternatives based on differences in the locations of the samples. It is a K-sample generalization of the Wilcoxon rank test.

An indication of the position of each sample is given by its average rank, calculated as the mean value of the ranks received by the sample individuals in the pooled sorted list. Let R_{ji}, $\quad 1 \leq j \leq K; \quad 1 \leq i \leq n_j$ be the rank (in the list of pooled ranks) of the i-th

element from the j-th sample. Then,

$$R_{j\cdot} = \frac{1}{n_j} \sum_{i=1}^{n_j} R_{ji} \tag{3.4}$$

is the average rank for the j-th sample and

$$R_{\cdot\cdot} = \frac{(R_{11} + \cdots + R_{1n_1}) + \cdots + (R_{K1} + \cdots + R_{Kn_K})}{N} = \frac{N+1}{2} \tag{3.5}$$

is the overall average.

The Kruskal-Wallis statistic is defined as

$$KW = \frac{12}{N(N+1)} \sum_{j=1}^{K} n_j \left(R_{j\cdot} - \frac{N+1}{2} \right)^2 \tag{3.6}$$

and H_0 is rejected for large values of KW. Its exact distribution under H_0 is determined, as for all the rank-based methods, by the general permutational distribution

$$\Pr(R_{11} = r_{11}, \ldots, R_{1n_1} = r_{1n_1}; \ldots; R_{K1} = r_{K1}, \ldots, R_{Kn_K} = r_{Kn_K}) = \frac{1}{N!}. \tag{3.7}$$

Asymptotically, the null distribution of KW is χ^2_{K-1} and this distribution is typically adequate for this statistic when either $K = 3$ and the three sample sizes are greater than 5 or when $K > 3$ and $n_j > 4$, for all j (Lehmann, 1975).

3.3.3 Normal Scores Test

An alternative for the Kruskal-Wallis test based on the normal scores of the ranks was proposed by Puri (1964). The basic idea is to transform the ranks to the expected values of the order statistics from a standard normal sample of size N.

Let the data from the K samples and the order statistics for the pooled sample be denoted by $X_{11}, \ldots, X_{1n_1}, \ldots, X_{K1}, \ldots X_{Kn_K}$ and $Z_{(1)}, \ldots, Z_{(N)}$, respectively. Assuming,

for a moment, that the random variables $\{Z_{(i)}\}$, $(1 \leq i \leq N)$ are independently normally distributed random variables, and that only their ranks, not the ralized values, are known, how could we reconstruct the sample values? $Z_{(s)}$ has rank s, and it is the s-th smallest value of a sample of size N from a standard normal distribution. A natural estimate for its expected value is:

$$a_N(s) = \mathrm{E}_{\Phi}(Z_{(s)}), \tag{3.8}$$

where Φ is the standard normal distribution. These expectations are known as *normal scores*. The values for equation (3.8), as a function of N and s may be calculated using subroutine `G01DAF` of the `NAg` library.

A two sample test that resembles the t-statistic, but depends only on the observed ranks of the pooled sample, has been studied by Klotz (1964). The resulting test rejects the null hypothesis $F_{X_1} = F_{X_2}$ when

$$T_{n_2} > c \tag{3.9}$$

where

$$T_{n_2} = \sum_{\alpha=1}^{n_2} a_N(R_{2\alpha}) \tag{3.10}$$

The distribution of this statistic is asymptotically normal. It has the property of having asymptotic efficiency relative to the t test greater than or equal to one for the *shifted model* case, $F_{X_1}(x) = F_{X_2}(x-\Delta)$, for all x, and for all continuous distributions (Lehmann, 1975); in this situation, this test is at least as powerful as the t test.

For a K sample univariate generalization, consider the following test statistic proposed by Puri (1964):

$$NS = \frac{N-1}{\sum_{i=1}^{N} [a_N(i)]^2} \sum_{j=1}^{K} \frac{1}{n_j} \left[\sum_{\alpha=1}^{n_j} a_N(R_{j\alpha}) \right]^2 \tag{3.11}$$

where the $a_N(i)$ were defined in equation (3.8). The null distribution of expression (3.11) tends to the $\chi^2_{(K-1)}$ distribution as the sample sizes tend to infinity. The χ^2-approximation works well even for relatively small total sample sizes (Lehmann, 1975).

3.3.4 Kiefer Tests

Kiefer (1959) constructed several K-sample analogues of the Kolmogorov-Smirnov test. This author, as Kolmogorov and Smirnov did, based his tests on some distance measures between the samples *edf*s . These tests have not been extensively used. This is probably due to the fact that their small sample properties, as well as their rates of convergence to the corresponding limiting distributions, remain unstudied. We consulted many papers whose references included Kiefer's paper. The majority of those works were connected with some theoretical results that Kiefer obtained in order to prove the limiting distributions of some of his statistics. We now discuss one of Kiefer's test statistics.

Kiefer's T statistic based on the *edf*s F_{X_j}, and may be written as

$$T = \sup_x \sum_{j=1}^{K} C_j[F_{X_j}(x) - \bar{F}_X(x)]^2 \tag{3.12}$$

where the C_js are positive constants and

$$\bar{F}_X(x) = \frac{1}{N} \sum_{j=1}^{K} n_j F_{X_j}(x) \tag{3.13}$$

is the sample *edf* of the pooled K samples.

Kiefer obtained the limiting distribution for T, with $C_j = n_j$, when the $n_j \to \infty$. Consider the function

$$A_h(a^2) = \frac{4}{a^h \Gamma\left(\frac{h}{2}\right)^{h/2}} \sum_{n=1}^{\infty} \frac{(\gamma_{(h-2)/2,n})^{h-2} \exp[-(\gamma_{(h-2)/2,n})^2 / 2a^2]}{[J_{h/2}(\gamma_{(h-2)/2,n})]^2} \tag{3.14}$$

where $h \geq 1$; $\Gamma(x)$ is the Gamma function ($x > 0$), $J_\nu(x)$ is the Bessel function of the

first kind, ($\nu \in \mathbb{R}$), and $\gamma_{\nu,n}$, ($n = 1, 2, \ldots$) are the positive zeros of J_ν. Now, if

$$\phi_K(x) = \begin{cases} A_K(x^2) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases},$$

then ϕ_{K-1}, is the limiting d.f. of the random variable $\sqrt{T}$. Kiefer (1959) tabulated this distribution for the values $K = 2, \ldots, 6$. No small sample tables are available for the null distribution of this test statistic, and, as Lehmann (1975) remarked, because of the slow convergence of the null distribution of the Smirnov test, it seems advisable to use the approximation ϕ_{K-1} only when the sample sizes are fairly large. For very small sample sizes, however, one could always calculate the exact permutational distribution over the *edf*.

3.3.5 Birnbaum and Hall Test

Some K-sample tests were developed by Birnbaum and Hall (1960) which are direct generalizations of the Kolmogorov-Smirnov two-sample test. They assume that the distributions of the K samples are continuous; the sample sizes may not be necessarily equal. The Birnbaum and Hall statistics are written as

$$D(n_1 \ldots, n_k) = \sup_{\substack{z,i,j \\ i \neq j}} [F_i(z) - F_j(z)]$$

and

$$D^+(n_1 \ldots, n_k) = \sup_{\substack{z,i,j \\ i<j}} [F_i(z) - F_j(z)].$$

Birnbaum and Hall (1960) obtained the null distribution functions for these statistics using a system of difference equation concerning the number of paths which exist from the origin of the K-dimensional unit cube and the point $(k_1 / n_1, \ldots, k_K / n_K)$, with $0 \leq k_i \leq n_i$. They also calculated the corresponding tables for two and three samples of equal sizes, for $2 \leq n_j \leq 40$. The derivation of the distribution functions is a straightforward matter using their system. However, very little else is known about

the distribution of the Birnbaum and Hall statistics or any other of their properties. Conover (1971, p. 320) mentioned an unproven conjecture of his own concerning the asymptotic null distribution for the three sample case with equal sample sizes. Although it is possible to obtain the distribution function for any case using Birnbaum and Hall's difference equations but the computations, particularly for unequal sample sizes, become rapidly too awkward to compensate the effort. In addition, we found only very few references to this paper in the literature consulted, and so we did not use it in further chapters.

3.3.6 Conover Tests

Conover (1965) discussed several K-sample Kolmogorov-Smirnov tests. All of them assume that the K samples are of equal size, say n, and that they come from a continuous distribution. Conover's approach was to reduce the K-sample case to different two-sample problems. This is achieved by first ranking the samples in order to use those ranks to select pairs of samples and then comparing the chosen samples pairwise. For univariate observations, this ranking is done by ordering the K samples in accordance with the extreme (minimum) value of each sample. In the multivariate case we would proceed analogously, ordering the K samples using their minimum ranks obtained with the diameter ranking of the 1-*MST*.

We now describe Conover's statistics and their exact null distributions. First, we order the samples within themselves according to their rankings. Let the ranks of the l-th ordered sample among the pooled sample's ranks be denoted by $Z_{1l} < Z_{2l} < \ldots < Z_{nl}$. Then Z_{il} is, in the univariate case, the i-th order statistic of the l-th sample; Z_{1l} will be referred to as the *extreme* of the l-th sample.

Next, we order the samples among themselves in accordances with their extremes. Let $Y_{11} < Y_{12} < \ldots < Y_{1K}$ denote the ordered extremes and let Y_{ij} be the i-th order statistic of the sample whose extreme is Y_{1j}.

Within this set up, Conover presented the following extensions of the Kolmogorov-Smirnov two sample test

$$D^{-}_{j_1,j_2}(K,n) = \sup_{\substack{y \\ j_1<j_2}} [S_{j_1}(y) - S_{j_2}(y)] \tag{3.15}$$

$$D^{+}_{j_1,j_2}(K,n) = \sup_{\substack{y \\ j_1<j_2}} [S_{j_2}(y) - S_{j_1}(y)] \tag{3.16}$$

where S_{j_k} is the *edf* of the j_k-th ordered sample. Several variations of these statistics might be proposed, always comparing the empirical distribution functions of pairs of samples, according to the alternative hypothesis of interest. For instance, to test the hypothesis of homogeneity against the alternative that at least one of the populations differs in location, one could use $D^{+}_{1,K}(K,n)$. To divide the K samples into groups having similar location parameters, one could use $D^{+}_{j,j+1}$, with $1 \leq j \leq K-1$. If the alternative hypothesis is that the populations differ by a scale parameter only, then $D^{-}_{1,K}(K,n)$ would be a suitable test statistic.

Conover obtained the following formulae for the distribution functions of $D^{-}_{j_1,j_2}(K,n)$ and $D^{+}_{j_1,j_2}(K,n)$

$$\Pr\left(D^{-}_{j_1,j_2}(K,n) \leq \frac{c}{n}\right) = 1 - \binom{K-j_1}{j_2-j_1}\binom{2n-2}{n+c} \Big/ \binom{K-j_1-(c+1)/n}{j_2-j_1}\binom{2n-2}{n-1}$$

$$+ \sum_{i=0}^{j_2-j_1-1} \frac{(n-1)_c\,(c+1)\,(K-j_1)_{j_2-j_1}\,(-1)^{j_2-j_1-i}\,(K-j_1-i-1)}{i!\,(j_2-j_1-1-i)!\,(nK-nj_1+n-1-ni)_c} \cdot (K-j_1-i)\,(nK-nj_1-c-1-ni) \tag{3.17}$$

$$\Pr\left(D^{+}_{j_1,j_2}(K,n) \leq \frac{c}{n}\right) = 1 - \binom{K-j_1}{j_2-j_1}\binom{2n-2}{n+c} \Big/ \binom{K-j_1+(c+1)/n}{j_2-j_1}\binom{2n-2}{n-1}$$

(3.18)

where $(n)_c = n(n-1)\cdots(n-c+1)$

Conover showed that both $D^-_{j_1,j_2}(K,n)$ and $D^+_{j_1,j_2}(K,n)$ have, asymptotically, the same distribution, and that it is independent of K. If we write $\lambda = \frac{c}{\sqrt{n}}$, then

$$\lim_{n\to\infty} \Pr\left(D^+_{j_1,j_2}(K,n) \le \frac{c}{n}\right) = \lim_{n\to\infty} \Pr\left(D^-_{j_1,j_2}(K,n) \le \frac{c}{n}\right) = 1 - e^{-\lambda^2} \qquad (3.19)$$

This approximation seems to work very well for sample sizes larger than 50.

3.3.7 Scholz-Stephens Test

The motivation of the paper by Scholz and Stephens (1987) was to extend the goodness of fit test statistic of Anderson and Darling (1954) to a K-sample nonparametric test of homogeneity against general alternatives.

The Anderson-Darling (1954) A^2_n goodness of fit statistic was proposed in order to test the hypothesis $F_X = F_0$, where F_0 is a specified distribution function, for a sample of size n. It can be written as:

$$A^2_n = n \int_{-\infty}^{\infty} \frac{\{F_n(x) - F_0(x)\}^2}{F_0(x)(1-F_0(x))}\, dF_0(x)$$

where $F_n(x)$ is the *edf* of the sample. A two-sample version of A^2_n is given by

$$A^2_{n_1 n_2} = \frac{n_1 n_2}{N} \int_{-\infty}^{\infty} \frac{\{F_{n_1}(x) - G_{n_2}(x)\}^2}{H_N(x)\,(1-H_N(x))}\, dH_N(x)$$

where

$$H_N(x) = \{n_1 F_{n_1}(x) + n_2 G_{n_2}(x)\} / N$$

is the pooled sample *edf*. This statistic is used to test the hypothesis $F = G$, without specifying the common distribution. Scholz and Stephens (1987) generalized this statistic for the K-sample case as:

$$A^2_{KN} = \sum_{j=1}^{K} n_j \int_{B_N} \frac{\{F_{jn_j}(x) - H_N(x)\}^2}{H_N(x)\{1 - H_N(x)\}} dH_n(x) \tag{3.20}$$

where F_{jn_j} and $H_N(x)$ denote the *edf* of the j-th sample and of the pooled sample, respectively, and $B_N = \{x \in \mathbb{R} | H_N(x) < 1\}$. If the pooled order sample is denoted by $Z_{(1)} < \cdots < Z_{(N)}$, then, assuming no ties, a computational formula for A^2_{KN} is given by

$$A^2_{KN} = \frac{1}{N} \sum_{j=1}^{K} \frac{1}{n_j} \sum_{i=1}^{N-1} \frac{(N M_{ji} - i n_j)^2}{j(n-j)} \tag{3.21}$$

where M_{ji} is the number of observations from the j-th sample that are not greater than $Z_{(i)}$. Under H_0, Scholz and Stephens proved that $\mathrm{E}\,(A^2_{KN}) = K - 1$ and

$$\sigma^2_{K,N} = \mathrm{var}\,(A^2_{KN}) = \frac{aN^3 + bN^2 + cN + d}{(N-1)(N-2)(N-3)} \tag{3.22}$$

where

$$a = (4g - 6)(K - 1) + (10 + 6g)H$$

$$b = (2g - 4)K^2 + 8hK + (2g - 14h - 4)H - 8h + 4g - 6$$

$$c = (6h + 2g - 2)K^2 + (4h - 4g + 6)K + (2h - 6)H + 4K$$

$$d = (2h + 6)K^2 - 4hK^2$$

$$g = \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \frac{1}{(N-i)j} \quad \xrightarrow[N \to \infty]{} \quad \int_0^1 \int_y^1 \frac{dx\,dy}{x(1-y)} = \frac{\pi^2}{6}$$

$$H = \sum_{j=1}^{K} \frac{1}{n_j}$$

$$h = \sum_{i=1}^{N-1} \frac{1}{i}.$$

The effect of the individual sample sizes is reflected through H, and is not negligible to order $1/N$. The null hypothesis is rejected for large values of A^2_{KN}.

Because of the definition of the M_{ji}s, it is plain that A^2_{KN} depends only on the ranks of the pooled sample. For small samples, Scholz and Stephens recommended estimating the exact significance level by the relative frequency $\hat{p}$ with which the realized value a^2_{KN} of A^2_{KN} is matched or exceeded when computing the test statistic for a large number M of random ranks permutations. $\hat{p}$ is an unbiased estimator of the true p value of the realized value of the test statistic and its variance can be controlled by the choice of M.

The asymptotic null distribution of A^2_{KN} is:

$$A^2_{K-1} \equiv \sum_{i=1}^{\infty} \frac{1}{i(i+1)} Y_i \tag{3.23}$$

where the $\{Y_i\}$s are independent random variables identically distributed χ^2_{K-1}. This distribution is the $(K-1)$-fold convolution of the limiting distribution of the Anderson-Darling statistic, A^2_n. For A^2_n, the approximation works remarkably well, even for very small sample sizes, and Scholz and Stephens pointed out that this also happens for the K-sample version. They also remarked, the formulae for the mean and the variance of A^2_{KN} are given in order to allow us to standarize the test statistic. The standarization was carried out as an attempt to remove any dependence due to the sample sizes that can affect the null distribution. This procedure worked satisfactorily, as some simulation experiments confirmed.

The standarized test statistic is written as

$$T_{KN} = \frac{A^2_{KN} - (K-1)}{\sigma_{K,N}} \tag{3.24}$$

thus the null hypothesis of homogeneity is rejected for large values of T_{KN}. The first four cumulants of expression (3.23) can be easily obtained, and, with them, Scholz and

Stephens fitted Pearson curves to this distribution and calculated asymptotic quantiles of the random variable $T_{K\infty} = t_m$, with $m = K-1$. It should be noted that as the number of samples increases, the distribution of t_m tends to the standard normal distribution. Scholz and Stephens (1987) gave some tables and expressions necessary to interpolate the quantiles of the null distribution for small sample sizes. These quantiles were reported to produce very accurate results even for sample sizes as small as $n_i = 5$, for $K \geq 3$.

For discrete distributions, or for when ties are present in the data, these authors gave two corrections to their statistic, based on changing the *edf* to the average of the left and the right limits of the ordinary *edf*. The same quantiles obtained for the continuous case can be applied for the modified statistic. These approximations produce very good results, as Scholz and Stephens observed in several simulation examples.

These authors also proved that this test is consistent for any alternative to H_0, provided that the ratio n_i / N is greater than 0 as $N \rightarrow \infty$ for each sample size n_i.

3.4 Puri and Sen Multivariate Rank Tests

Puri and Sen (1971:§5) presented a class of linear rank order statistics which is asymptotically equivalent to the Lawley-Hotelling's generalized T^2 statistic.

Three examples of the tests constructed using Puri and Sen's method are the multivariate multisample ranks sum test (*MMRST*), the multivariate multisample median test (*MMMT*) and the multivariate multisample normal scores test (*MMNST*).

The *MMRST* statistic is based on the differences between the samples' average ranks and the average rank of the pooled sample. The *MMMT* statistic uses the differences in the proportion of individuals with values less than or equal to the median for each sample and the corresponding proportion for the pooled sample. Finally, the *MMNST* works with linear combinations of the normal scores values calculated for the ranks obtained from the pooled observations on each variable which correspond to the observations in every sample The weights in the linear combinations for these functions of ranks in the Puri and Sen general statistic are given by the inverse of the

dispersion matrix corresponding to the values of the functions of the ranks.

The two first statistics mentioned are analogous to the corresponding likelihood ratio tests and are based on a quadratic form of the differences between the vectors of mean ranks of each sample and the vector of mean ranks for the pooled sample. The *MMNST* has been briefly mentioned by Friedman and Rafsky (1979) as a competitor for the parametric test when the observations come from a non normal distribution.

Puri and Sen (1971) showed that their class of statistics, asymptotically have a χ^2 distribution with $p\,(K-1)$ degrees of freedom and are equivalent to the likelihood ratio test based on Hotelling's T^2 statistic. We now review their general procedure and then describe the three tests based on it that will be used in further chapters.

3.4.1 Permutation Rank Order Tests

We begin by stating the basic rank permutation principle. This is used to obtain tests which depend only on the observed ranks of the pooled observations from the K samples for each of the p variates.

For each i, let us rank the N pooled observations from the i-th variable $X_{i\alpha}^{(j)}$, $i = 1, \ldots, p$; $\alpha = 1, \ldots, n_j$; $j = 1, \ldots, K$, in ascending order of magnitude, and denote the rank $X_{i\alpha}^{(j)}$, by $R_{i\alpha}^{(j)}$. Then, the p-variate vector $\mathbf{X}_{\alpha}^{(j)}$ produces the rank vector

$$\mathbf{R}_{\alpha}^{(j)} = (R_{1\alpha}^{(j)}, \ldots, R_{p\alpha}^{(j)}) \quad \alpha = 1, \ldots, n_j,\ j = 1, \ldots, K$$

The N rank vectors arising from the N pooled observations can be put together in the $(p \times N)$ rank matrix

$$\mathbf{R}_N = \begin{pmatrix} R_{11}^{(1)} & \ldots & R_{1n_1}^{(1)} & \ldots & R_{11}^{(K)} & \ldots & R_{1n_K}^{(K)} \\ . & & . & & . & . & \\ . & & . & & . & . & \\ . & & . & & . & . & \\ R_{p1}^{(1)} & \ldots & R_{pn_1}^{(1)} & \ldots & R_{p1}^{(K)} & \ldots & R_{pn_K}^{(K)} \end{pmatrix} \tag{3.25}$$

Each row of this matrix is a permutation of the integers $1, 2, \ldots, N$ and thus $\mathbf{R}_N$ is a random matrix which can have $(N!)^p$ possible realizations. We say that two rank matrices are *permutationally equivalent* if one can be obtained from the other by a rearrangement of its columns. Then, $\mathbf{R}_N$ is permutationally equivalent to the random matrix

$$\mathbf{R}_N^* = \begin{pmatrix} 1 & 2 & \cdots & N \\ R_{21}^* & R_{22}^* & \cdots & R_{2N}^* \\ \cdots & \cdots & \cdots & \cdots \\ R_{p1}^* & R_{p2}^* & \cdots & R_{pN}^* \end{pmatrix} \tag{3.26}$$

$\mathbf{R}_N^*$ contains the elements of $\mathbf{R}_N$ permuted within each column, and can have $(N!)^{p-1}$ possible realizations. As the p variates $X_{i\alpha}^k$ $i = 1, \ldots, p$, are, in general, stochastically dependent, the joint distribution of the elements of $\mathbf{R}_N$ (or $\mathbf{R}_N^*$) will depend on the unknown distribution $F \in \mathcal{F}$, even under the hypothesis of homogeneity. Let $\mathcal{R}_N^*$ denote the set of all the possible realizations of $\mathbf{R}_N^*$; then the unconditional distribution of $\mathbf{R}_N^*$ over this set depends on the distributions $F_1, \ldots F_K$, even under the null hypothesis. However, if H_0 holds, then the vectors $\mathbf{X}_\alpha^{(j)}$, $\alpha = 1, \ldots n_j, j = 1, \ldots, K$ are independent and identically distributed random vectors. Hence, their joint distribution is invariant under any permutation of the vectors. This implies that the conditional distribution of $\mathbf{R}_N$ over the set of $N!$ possible permutations on $S(\mathbf{R}_N^*)$, which can be obtained by all the permutations of the columns of $\mathbf{R}_N^*$ is uniform under H_0.
This can be written as:

$$\Pr\{\, \mathbf{R}_N = \mathbf{r}_N \,|\, S(\mathbf{R}_N^*), H_0 \,\} = \frac{1}{N!}, \tag{3.27}$$

for all $\mathbf{r}_N \in S(\mathbf{R}_N^*)$. This null distribution is independent of the parent *cdf* $F \in \mathcal{F}$.
Next, for each entry in the rank matrix, we associate a rank score value, defined by

$$E_{N,\alpha}^{(i)} = J_{N(i)}\left(\frac{\alpha}{N+1}\right) \quad 1 \le \alpha \le N, \qquad i = 1, \ldots, p. \tag{3.28}$$

where $J_{N(i)}$ is a function defined at the points $\frac{1}{N+1}, \ldots, \frac{N}{N+1}$.
The form of $J_{N(i)}$ would depend on the class of alternative we are interested in. For location alternatives, two choices would be

$$J_{N(i)}\left(\frac{\alpha}{N+1}\right) = \frac{\alpha}{N+1}$$

or taking $E_{N\alpha}^{(i)}$ as the expected value of the α-th smallest observation of a sample of a size N from a standard normal distribution.
For the scale problem, two plausible functions are

$$J_{N(i)}\left(\frac{\alpha}{N+1}\right) = \begin{cases} 1 & \text{if } |\alpha - (N+1)/2| \ge b \\ 0 & \text{otherwise} \end{cases}$$

$$J_{N(i)}\left(\frac{\alpha}{N+1}\right) = \left(\tfrac{\alpha}{N+1} - \tfrac{1}{2}\right)^2.$$

where $b = \frac{1}{2}N + 1$.
Another possibility for the scale problem is to make $E_{N\alpha}^{(i)}$ equal to the square of the value of the corresponding normal score.

Replacing the ranks $R_{i\alpha}^{(j)}$ in $\mathbf{R}_N$ by $E_{N,R_{i\alpha}^{(j)}}^{(i)}$ for all $i = 1, \ldots p$, $\alpha = 1, \ldots, n_j$ and $j = 1, \ldots K$, we obtain the corresponding $(p \times N)$ matrix $\mathbf{E}_N$ of general scores:

$$\mathbf{E}_N = \begin{pmatrix} E^{(1)}_{NR^{(1)}_{11}} & \cdots & E^{(1)}_{NR^{(1)}_{1n_1}} & \cdots & E^{(1)}_{NR^{(K)}_{11}} & \cdots & E^{(1)}_{NR^{(K)}_{1n_K}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ E^{(p)}_{NR^{(1)}_{p1}} & \cdots & E^{(1)}_{NR^{(p)}_{pn_1}} & \cdots & E^{(p)}_{NR^{(K)}_{p1}} & \cdots & E^{(p)}_{NR^{(K)}_{pn_K}} \end{pmatrix} \tag{3.29}$$

The average rank scores for variate i of the K samples is defined as

$$T^{(j)}_{Ni} = \frac{1}{n_j} \sum_{\alpha=1}^{n_j} E^{(i)}_{NR^{(j)}_{i\alpha}}, \qquad j = 1, \ldots K, \quad i = 1, \ldots, p$$

and the pooled rank average can be written as

$$\bar{E}^{(i)}_N = \sum_{\alpha=1}^{N} E^{(i)}_{N,\alpha} / N, \qquad i = 1 \ldots p$$

We have that, under the null hypothesis,

$$\operatorname{cov}(T^{(j)}_{Nh}, T^{(k)}_{Ni}) = \frac{(N - \delta_{jk})}{n_j (N-1)} v_{hi}(\mathbf{R}^*_N)$$

where δ_{jk} is the Kronecker delta, and

$$v_{hi}(\mathbf{R}^*_N) = \frac{1}{N} \sum_{j=1}^{K} \sum_{\alpha=1}^{n_j} E^{(h)}_{NR^{(j)}_{h\alpha}} E^{(i)}_{NR^{(k)}_{i\alpha}} - \bar{E}^{(h)}_N \bar{E}^{(i)}_N$$

Puri and Sen (1971, §5.4) proposed the following statistic which is an analogue of the Lawley-Hotelling's generalized T^2:

$$\mathcal{L}_N = \sum_{j=1}^{K} n_j \left[(\mathbf{T}^{(j)}_N - \mathbf{E}_N) \mathbf{V}^{-1}(\mathbf{R}^*_N) (\mathbf{T}^{(j)}_N - \mathbf{E}_N)' \right]. \tag{3.30}$$

They also proved that, under very general conditions for the functions $J_{N(i)}$ $i = 1, \ldots, p$, the permutation distribution of $\mathcal{L}_N$ is, asymptotically, $\chi^2_{p(K-1)}$.

The final subsections of this chapter describe three multivariate multisample tests which are particular cases of the statistic (3.30)

3.4.2 Multivariate Multisample Ranks Sum Test

The *MMRST* statistic is given by

$$\mathcal{L}_{MMRST} = \sum_{j=1}^{K} n_j (\mathbf{r}_j - \bar{\mathbf{r}})' \mathbf{V}^{-1} (\mathbf{r}_j - \bar{\mathbf{r}}) \tag{3.31}$$

where $\mathbf{r}_j$ is the vector (of length p) of average ranks for the j-th sample, $\bar{\mathbf{r}}$ is the vector of average ranks for the pooled sample and $\mathbf{V}$ is the dispersion matrix for the vector of ranks. The i-th element of $\mathbf{r}_j$ is

$$r_{ji} = \frac{1}{n_j} \sum_{\alpha=1}^{n_j} r_{ji\alpha} \tag{3.32}$$

where $r_{ji\alpha}$ is the rank of the i-th variable for the α-th individual from the j-th sample. The i-th element of $\bar{\mathbf{r}}$ is

$$\bar{\mathbf{r}}_i = \frac{1}{N} \sum_{j=1}^{K} \sum_{\alpha=1}^{n_i} r_{ji\alpha} \tag{3.33}$$

The (i, l) element of the variance-covariance matrix $\mathbf{V}$ is given by

$$v_{il} = \frac{1}{N} \sum_{j=1}^{K} \sum_{\alpha=1}^{n_j} r_{ji\alpha} r_{jl\alpha} - \bar{\mathbf{r}}_i \bar{\mathbf{r}}_l \tag{3.34}$$

3.4.3 Multivariate Multisample Median Test

For the *MMMT* statistic, the necessary definitions are analogous to those presented for the *MMRST*. The *MMMT* statistic is given by

$$\mathcal{L}_{MMMT} = \sum_{j=1}^{K} n_j (\mathbf{p}_j - \bar{\mathbf{p}})' \mathbf{V}_0^{-1} (\mathbf{p}_j - \bar{\mathbf{p}}). \tag{3.35}$$

where $\mathbf{p}_j$ is the vector (of length p) of proportions from the j-th sample which are less than or equal to the median of the pooled sample; its i-th element is expressed as

$$\mathbf{p}_{ji} = \frac{1}{n_j}\sum_{\alpha=1}^{n_j} x_{ji\alpha}, \quad \text{where} \quad x_{ji\alpha} = \begin{cases} 1 & \text{if } r_{ji\alpha} \leq \frac{1}{2}N \\ 0 & \text{otherwise} \end{cases}.$$

$\bar{\mathbf{p}}$ is the vector of proportions of observations from the pooled sample that less than or equal to the pooled sample median; its i-th element is

$$\bar{\mathbf{p}}_i = \frac{1}{N}\sum_{j=1}^{K}\sum_{\alpha=1}^{n_j} x_{ji\alpha} \tag{3.36}$$

Finally, the (i, l) element of the dispersion matrix $\mathbf{V}_0$ is given by

$$v_{il}^{(0)} = \frac{1}{N}\sum_{j=1}^{K}\sum_{\alpha=1}^{n_j} x_{ji\alpha}x_{jl\alpha} - \bar{\mathbf{p}}_i\,\bar{\mathbf{p}}_l \tag{3.37}$$

A FORTRAN subroutine, published by Schwertman (1982), is available to calculate the *MMRST* and *MMMT*statistics.

3.4.4 Multivariate Multisample Normal Scores Test

Using the normal scores of each variable, as calculated in Subsection 3.3.3, in order to define the functions $J_{N(i)}$ appearing in equation (3.28), the *MMNST* statistic may be written as

$$\mathcal{L}_{MMNST} = \sum_{j=1}^{K} n_j\,(\mathbf{s}_j - \bar{\mathbf{s}})'\,\mathbf{V}_1^{-1}\,(\mathbf{s}_j - \bar{\mathbf{s}}) \tag{3.38}$$

where $\mathbf{s}_j$ is the vector (of length p) of averages for the normal scores calculated for the pooled sample corresponding to observations from the j-th sample, that is

$$\mathbf{s}_{ji} = \frac{1}{n_j}\sum_{\alpha=1}^{n_j} s_{ji\alpha}$$

where $s_{ji\alpha}$ is the normal score corresponding to the α-th individual on the i-th variable of the j-th sample; their average is

$$\bar{\mathbf{s}}_i = \frac{1}{N} \sum_{j=1}^{K} \sum_{\alpha=1}^{n_j} s_{ji\alpha}. \tag{3.39}$$

The average of the normal scores over the K samples and the dispersion matrix $\mathbf{V}_1$ from equation (3.38) are calculated exactly as for the other two multivariate multisample rank tests described before. Thus, the (i, l) element of the dispersion matrix for $\mathcal{L}_{MMNST}$, $\mathbf{V}_1$, is given by

$$v_{il}^{(1)} = \frac{1}{N} \sum_{j=1}^{K} \sum_{\alpha=1}^{n_j} s_{ji\alpha} s_{jl\alpha} - \bar{\mathbf{s}}_i \bar{\mathbf{s}}_l \tag{3.40}$$

We include these tests in our work as they are the most common nonparametric procedures used in multivariate analysis. As we saw, they require that the observations can be ranked within each variable and so their application is restricted to a class of data smaller than for the multivariate runs tests.

Chapter 4

Tests Based on Contingency Tables

4.1 Introduction

Friedman and Rafsky (1979) described another class of tests based on dividing the nodes of the 1-*MST* into two mutually exclusive categories using any criterion independent of the sample labelling. In the two sample case, they suggested considering the degree sequence of the 1-*MST* as the partition criterion. If n_1 and n_2 are the sample sizes, and d_1 and d_2 are the number of nodes with degree 1 and degree greater than 1 in the spanning tree, it is possible to form the following contingency table

	X_1	X_2	
deg = 1	O_{11}	O_{12}	d_1
deg ≥ 2	O_{21}	O_{22}	d_2
	n_1	n_2	N

As we are conditioning on the observed spanning tree, the row and column totals of the table can be regarded as fixed and so, under the null hypothesis of homogeneity, O_{11} follows the usual hypergeometric distribution, so

$$\Pr[O_{11} = o_1] = \frac{\binom{d_1}{o_1}\binom{d_2}{o_2}}{\binom{N}{n_1}} \tag{4.1}$$

where $o_2 = n_1 - o_1$.

Friedman and Rafsky's motivation for using the degree 1 nodes as a partitioning criterion was the supposition that these nodes 'tend to be found at the edges of the point scatter so that one might expect this test to be sensitive to scale alternatives'. We call degree 1 nodes the *leaves* of the graph. However, the results presented by Steele et al. (1987) cast some doubts on the validity of this assumption. These authors proved that, if $V_{k,n}$ is the number of nodes of degree k in the 1-*MST* generated by n points independently and identically distributed with a density f in $\mathbb{R}^p$, then, with probability 1

$$\lim_{n\to\infty} n^{-1} V_{k,n} = \alpha_{k,p} \qquad \text{for all} \quad k \geq 1, \text{ and } p \geq 2$$

for some positive constant $\alpha_{k,p}$. As Steele et al. (1987) noted, it seems impossible to try to find an analytic approach that will determine the values of the coefficients $\alpha_{k,p}$. One has to rely on estimating them via computer simulation or on establishing bounds for these constants. The outcome of some simulations performed by Steele et al. (1987) indicates that the limit for the number of leaves for the 1-*MST* in two dimensions, $\alpha_{1,2}$, is near 2/9, and not 0, as one might think, in accordance with the one-dimensional case, where the 1-*MST* coincides with the sorted list of observations. Steele et al.'s calculations show that there is a substantial amount of leaves located relatively near the centre of the observations, at least for 1-*MST*s in the plane. Nothing else is known about the values of $\alpha_{k,p}$, for $p > 2$.

However, Friedman and Rafsky (1979) gave evidence that a two-sample test based on the number of nodes with degree 1 from the first sample achieved quite good power results, particularly for higher dimensional data ($p \geq 10$) and scale alternatives. We considered that it would be worthwhile to study a K-sample generalization of this test, and we do this in the next section.

Later in this chapter, we discuss a technique outlined by Robinson (1987) in order to highlight differences among pairs of samples. The idea is to examine the number of edges in the graph defined by points from different pairs of samples. The observed frequencies of such edges for every pair of samples may be used to construct a test analogous to a χ^2 goodness of fit test.

4.2 Degree 1 Tests

4.2.1 Extensions to the K-sample Case

Extending the degree 1 test of Friedman and Rafsky to the K-sample situation is a straightforward matter. Consider the $2 \times K$ contingency table

	X_1	X_2		X_K	
deg 1	O_{11}	O_{12}	...	O_{1K}	d_1
deg ≥ 2	O_{21}	O_{22}	...	O_{2K}	d_2
	n_1	n_2		n_K	N

Under H_0, the expected frequencies for each cell may be written as $E_{ij} = \frac{d_i n_j}{N}$, and, asymptotically, the statistic

$$T_1 = \sum_{i=1}^{2} \sum_{j=1}^{K} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.2}$$

follows a χ^2 distribution with $(K-1)$ degrees of freedom.

We also applied this test using the degree 1 nodes of the 1-*RNG*. As we mentioned in Chapter 1, the number of leaves for this graph is much less than such number for the 1-*MST*. Because the 1-*RNG* has an edge density greater than or equal to that corresponding to the 1-*MST*, we would expect its leaves to be located more towards the edges of the point cloud than for the 1-*MST*. Thus, a test based on the leaf counts for the 1-*RNG* should perform well under scale alternatives, provided that we observe a large enough number of leaves. One problem with this test is that it is not possible to apply the χ^2 approximation for T_1 if there are samples which have very few degree 1 nodes.

We exclude higher order minimum spanning trees or relative neighbourhood graphs as the number of leaves would be too small to convey enough information about the differences between the samples. On the other hand, we did not consider the frequencies of degree 1 nodes for graphs as sparse as 1-*NNG* or some of Urquhart's relative neighbourhood graphs because the leaves of these graphs might correspond

mainly to extremes of fragments rather than to points located in the peripheries of the pooled sample.

4.2.2 Differences in the Number of Leaves

As a way of gaining some knowledge about the variation of the number of leaves we performed some simulations. Some relating results have been already mentioned in Chapter 1. We now study the proportions of leaves in the 1-*MST* and 1-*RNG* for different multivariate distributions. The data were generated using p independent random variables identically distributed as Uniform (U), Cauchy (C), $t(2)$ (t) Normal (N), Lognormal (L) and Exponential (E); the number of nodes considered were 20, 50, 100, 200, and 500. For every combination of N and p, we simulated 100 samples. The results presented correspond to the median of the number of leaves in those samples. Figure 4-1 shows the proportion of leaves found in the 1-*MST* and the 1-*RNG* for the above numbers of nodes in 2 and 20 dimensions. For the bivariate case, the number of degree 1 nodes in the 1-*MST* seems to tend rather quickly to $\alpha_{1,2}$. The convergence to $\alpha_{1,p}$ becomes slower for higher dimensions, as it is apparent from the larger spread of the number of leaves observed in these cases.

This fact is illustrated in the r.h.s. part of Figure 4-1. It is possible to spot a pattern concerning the different distributions. For all the dimensions considered, the number of leaves for the 1-*MST* induces the following ordering in the distributions:

C, L, E, t, U, N,

while the corresponding ordering for the 1-*RNG* is

U, N, E, L, t, C.

It is clear that tests based on the 1-*RNG* may not be generally adequate due to the small proportion of leaves usually observed for this graph.

For symmetric distributions, the 1-*MST* yields a larger proportion of leaves, while this does not seems to happen for the 1-*RNG*. It can be seen, from Figure 4-1 that the convergence to α_{1p} is much slower for higher dimensional configurations.

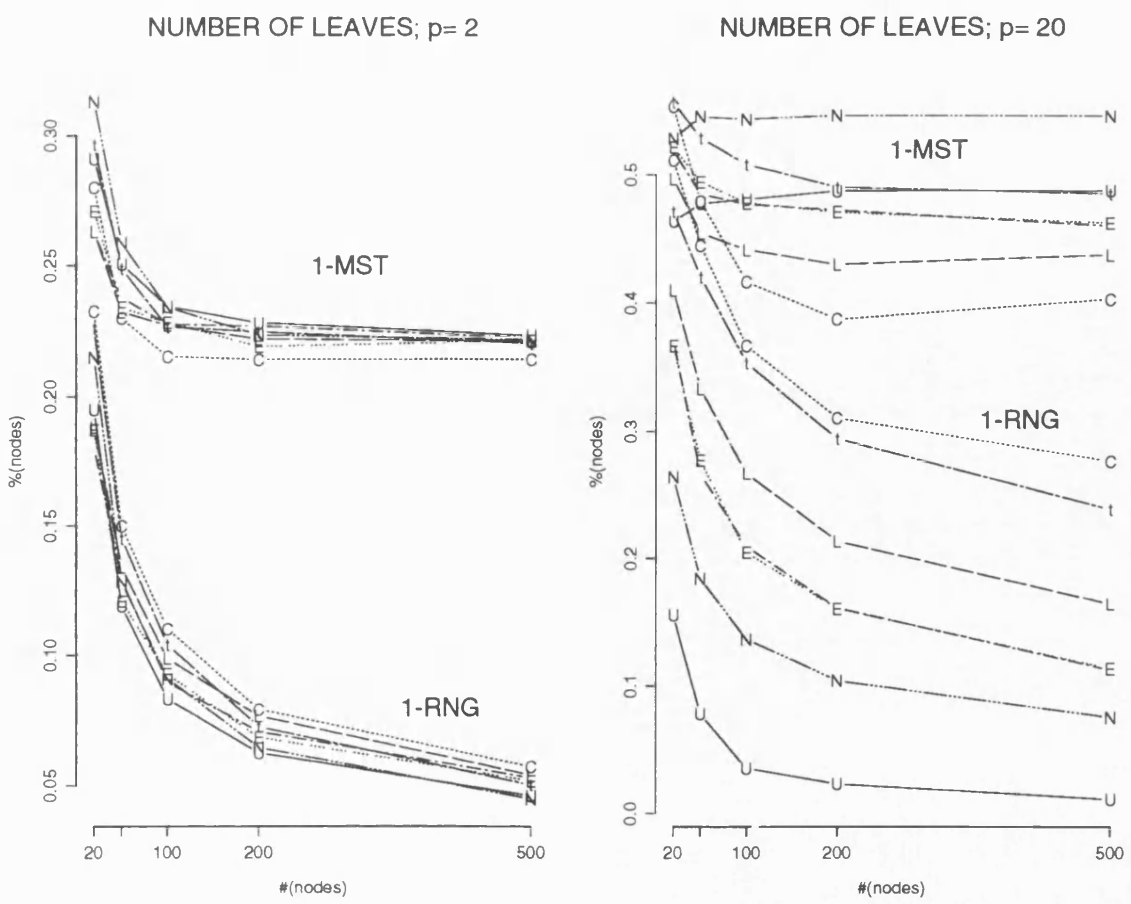


Figure 4-1: Number of leaves for fixed dimensions

Figure 4-2 shows the variation of the number of leaves for two fixed values of the number of nodes. Similar orderings to those mentioned may be observed in that figure. To explore with more detail how the shape of the underlying multivariate distribution affects the number of leaves, we may use the three following multivariate indices. First, the *total variation*, defined as

$$TV = \mathrm{tr}\,(\mathbf{S}) \tag{4.3}$$

where **S** is the variance covariance matrix.

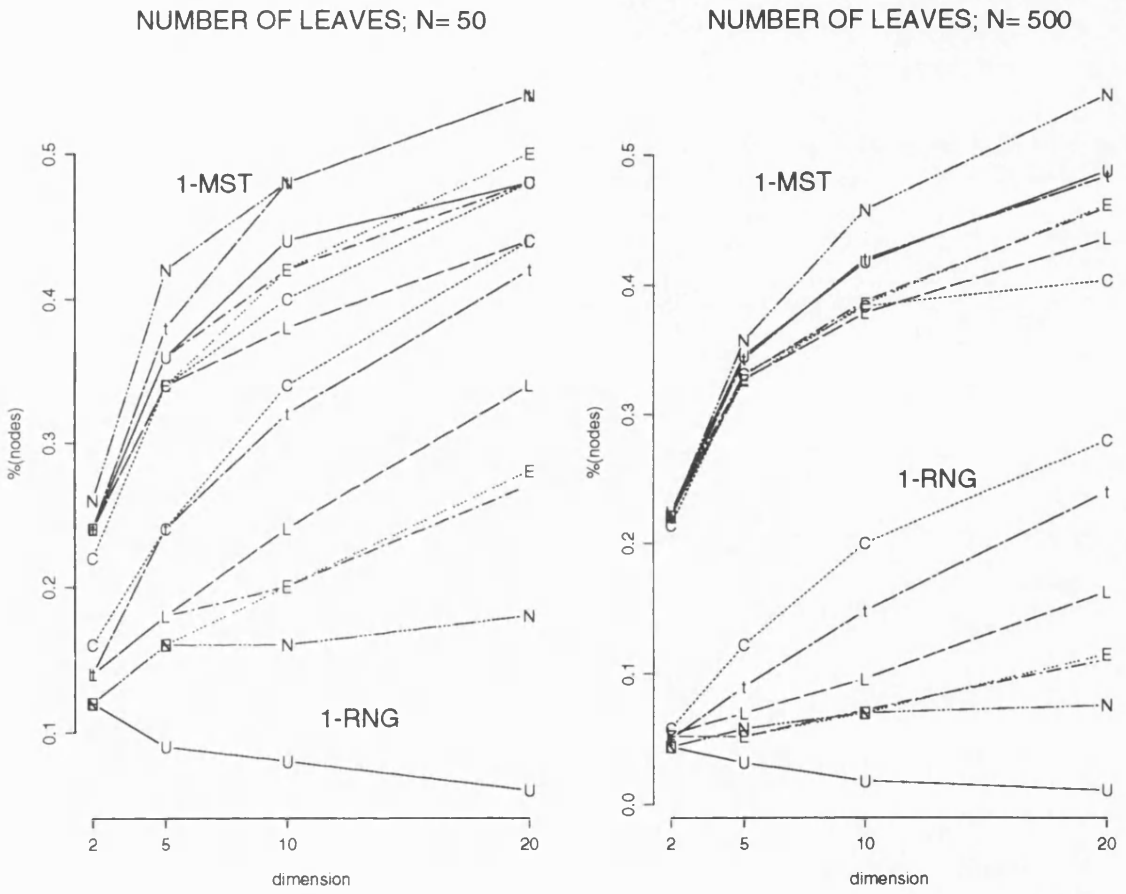


Figure 4-2: Number of leaves for fixed number of nodes

Further insight might be obtained from the measures of multivariate skewness and kurtosis proposed by Mardia (1970)

$$b(1,p) = \frac{1}{N^2}\sum_{r,s}^{N} g_{rs}^3 \tag{4.4}$$

$$b(2,p) = \frac{1}{N}\sum_{r=1}^{N} g_{rr}^2 \tag{4.5}$$

where

$$g_{rs} = (\mathbf{x}_r - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_s - \bar{\mathbf{x}})$$

Mardia (1970) showed that the asymptotic distributions of $b(1,p)$ and $b(2,p)$, under the null hypothesis that the underlying population is multivariate normal, are

$$Ib_1 = \frac{n}{6} b(1,p) \approx \chi^2_{df} \qquad \text{where} \qquad df = \frac{1}{6} p(p+1)(p+2) \tag{4.6}$$

$$Ib_2 = \frac{b(2,p) - p(p+2)}{[8p(p+2)]^{1/2}} \approx N(0,1) \tag{4.7}$$

Mardia (1970) also gave some approximate distributions for small sample sizes.

We don't show it here, but it can be seen that there is a relation between orderings mentioned before and these measures. Broadly speaking, the more "compact" (i.e. with small tr(**S**)) symmetric and flat a distribution is, the more and less leaves it will have in 1-*MST* and 1-*RNG*, respectively. This should be taken into account when applying these tests to a particular data set.

Although Steele et al.'s result is valid for the 1-*MST* obtained for points from any distribution, the convergence rate to $\alpha_{1,p}$ slows down as p increases, and so, even for patterns with a moderately large number of points the proportion of leaves varies considerably for different densities.

We also attempted to evaluate to what extent we should expect the leaves of a graph to be located in places corresponding to the outer layers of the data convex hull.

To do so, we first calculated the *mediancentre* of the observations, using Bedall and Zimmermann's (1979) algorithm. The mediancentre is an analogue of the univariate median, and thus, is a robust estimate of the centre of a distribution. It may be defined, for a set S of n points in $\mathbb{R}^p$, as the point m such that

$$f(m) = \sum_{x \in S} \| x - m \|_2 \leq \sum_{x \in S} \| x - y \|_2 = f(y) \qquad \text{for all } y \in \mathbb{R}^p$$

where $\| \cdot \|_2$ is the Euclidean distance. The mean of S would be obtained similarly by considering the sum of squares of the distances in the previous expressions. Some

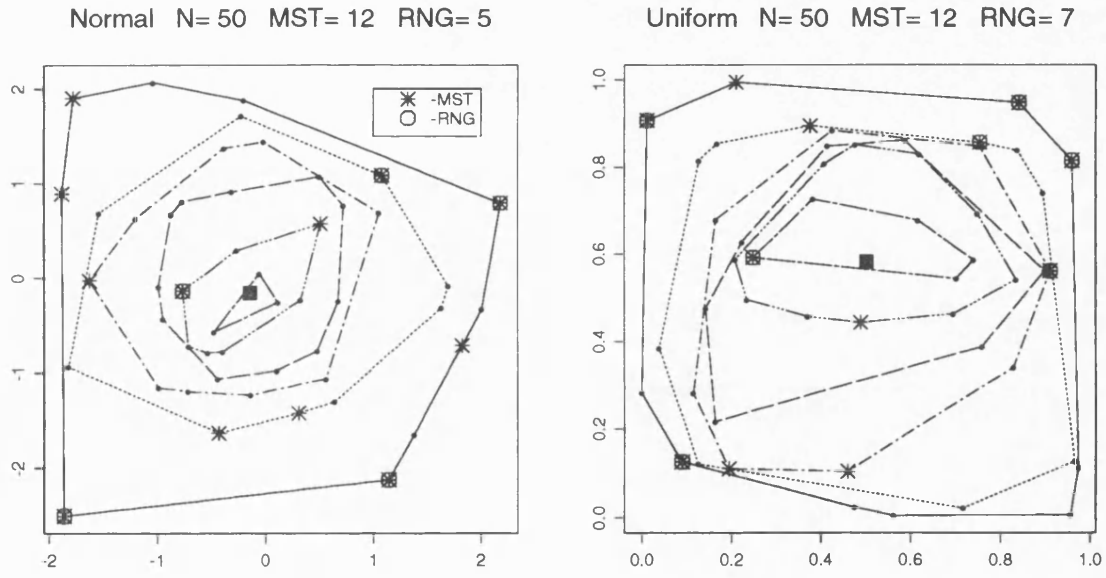


Figure 4-3: Degree 1 nodes; Normal and Uniform distributions

alternatives to the mediancentre as a measure of multivariate location have been studied by Green (1981). A recent survey of multivariate medians has been compiled by Small (1990).

Four examples for bivariate distributions appear in Figures 4-3 and 4-4. The mediancentre is indicated in those figures by a black square. The number of leaves for the 1-*MST* is close to $2N/9$, in accordance with a conjecture of Steele et al. (1987), even for the small values of N that we considered for the examples. These figures illustrate the fact that, for bivariate distributions, although the majority of the leaves

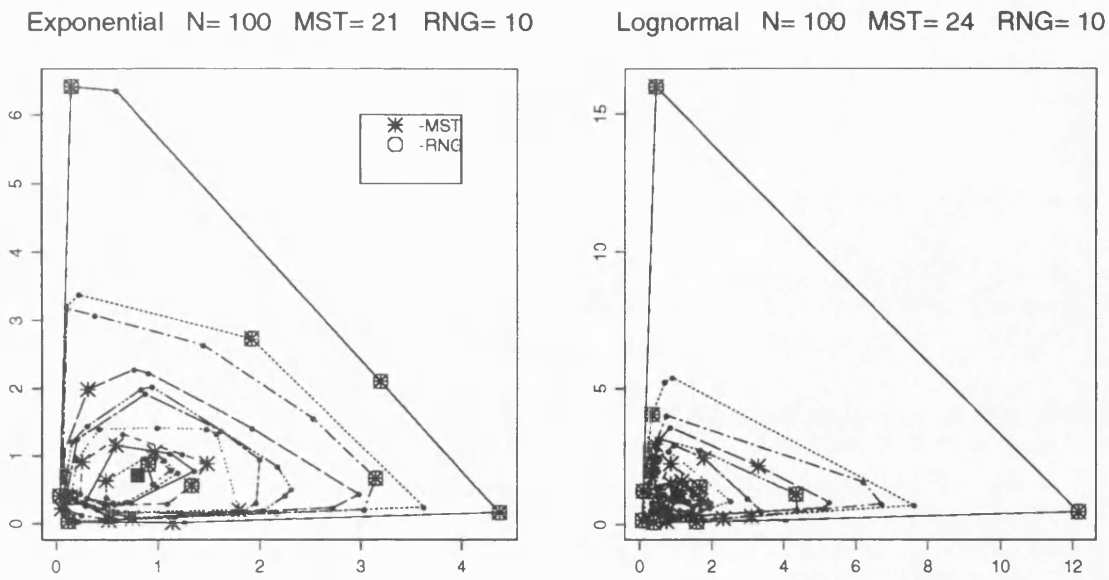


Figure 4-4: Degree 1 nodes; Exponential and Lognormal distributions

appear in the outer layers of the convex hull sequence, it is possible to find a number of them quite near to the centre of the data.

In order to examine the variations of the proportion of leaves located in the extremes of the data, we compared the ranks of the distances to the mediancentre for leaves and non-leaves using a one-sided Wilcoxon test (Lehmann, 1975). This nonparametric two-sample test was used as the distribution of the distances is usually far from being Normal; it can be written as

$$W_{XY} = W_X - \frac{1}{2}\, n_1\,(n_1 + 1),$$

where W_X is the sum of the ranks from the pooled sorted observation list corresponding to the first sample.

It can be shown that $\mathrm{E}(W_{XY}) = \frac{1}{2}\, n_1\, n_2$ and $\mathrm{var}(W_{XY}) = \frac{1}{12} n_1\, n_2\,(N+1)$, and that the asymptotic distribution of the quantity

$$W^*_{XY} = \frac{W_{XY} - \mathrm{E}(W_{XY})}{\sqrt{\mathrm{var}(W_{XY})}}$$

is standard normal.

The alternative hypothesis is that the ranks of the distances to the mediancentre corresponding to the degree 1 nodes were larger than those for nodes with higher degrees. We generated 100 samples of several combinations of dimension, number of nodes and distribution and calculated the Wilcoxon statistic for the ranks of the distances to the mediancentre. Table 4.1 presents the results for those combinations. This table shows that these tests should perform better for higher dimensional data as the number of leaves and the number of times in which such nodes are located in the periphery of the data increases with the number of dimensions for all the distributions for both, the 1-*MST* and the 1-*RNG*.

$p = 2$

	1-*MST*				1-*RNG*			
	$N = 20$		$N = 100$		$N = 20$		$N = 100$	
	signif	deg 1	signif	deg 1	signif	deg 1	signif	deg 1
U	4	6	1	23	18	3	11	8
C	81	6	39	22	81	5	97	10
χ^2_2	32	6	22	22.5	41	4	80	8
L	35	5	7	22	56	4	79	10
N	72	6	93	24	76	4	95	9
t (2)	78	6	78	23	80	4	96	10
E	28	5	20	22	38	4	71	9

$p = 5$

	1-*MST*				1-*RNG*			
	$N = 20$		$N = 100$		$N = 20$		$N = 100$	
	signif	deg 1	signif	deg 1	signif	deg 1	signif	deg 1
U	13	8	22	36	22	3	21	7
C	94	8	94	32	94	7	100	20
χ^2_2	30	8	33	33	54	5	89	11
L	47	7	19	33	58	6	91	14
N	88	9	100	39	62	5	100	12
t (2)	92	9	100	36	89	7	100	18
E	38	8	30	34	52	5	86	12

$p = 20$

	1-*MST*				1-*RNG*			
	$N = 20$		$N = 100$		$N = 20$		$N = 100$	
	signif	deg 1	signif	deg 1	signif	deg 1	signif	deg 1
U	21	9	38	48	15	2	13	3
C	99	11	100	42	98	11	100	37
χ^2_2	35	10	87	48	47	7	95	22
L	74	10	81	10	96	4	99	27
N	93	11	100	54.5	79	5	100	13.5
t (2)	99	11	100	51	97	9	100	35
E	40	10	84	48	41	7	97	21

signif = $\#(W^*_{XY} > z_{0.95})$;
deg 1 = median of #(degree 1 node)

Table 4.1: 100 samples; distances to the mediancentre for degree 1 nodes

4.3 Sample Pairwise Comparisons

Robinson (1987) suggested the use of the number of links between each possible combination of sample identities as a follow-up analysis for the K-sample multivariate runs test. Let C_{jk} be the observed number of edges in $\mathcal{G}_X$ defined by points from samples j and k, with $j \leq k$; these counts may be compared with their expected values calculated under H_0. If, for a particular pair of different samples, the observed count is much smaller than E_{jk}, then we would have an indication of existing differences in the distributions of the respective populations.

The expected values of C_{jk} under the hypothesis of homogeneity were obtained by Robinson (1987) in the following way. As the probability that any edge of the 1-*MST* links points from samples j and k is

$$P_{jk} = \begin{cases} \dfrac{2n_j n_k}{N(N-1)} & \text{if} \quad j \neq k \\ \\ \dfrac{n_j(n_j-1)}{N(N-1)} & \text{if} \quad j = k \end{cases}, \tag{4.8}$$

then the expected number of edges of the 1-*MST* joining points with sample labels j and k, say, E_{jk}, is obtained by summing over the edges of the 1-*MST* and is given by

$$E_{jk} = \begin{cases} \dfrac{2n_j n_k}{N} & \text{if} \quad j \neq k \\ \\ \dfrac{n_j(n_j-1)}{N} & \text{if} \quad j = k \end{cases}. \tag{4.9}$$

In general, if the spanning graph $\mathcal{G}_X$ used to link the observations has m links, then E_{jk} can be written as

$$E_{jk} = m P_{jk}. \tag{4.10}$$

Robinson (1987) also mentioned the possibility of considering a χ^2-type test using these expressions. We can construct the following table

sample pairs	$(1, K)$	...	$(K-1, K)$	
# (edges $\in \mathcal{G}_X$)	C_{1K}	...	$C_{(K-1)K}$	m

where the E_{jk}s are the corresponding expected values, obtained using expression (4.10). This table is a useful aid for suggesting hypotheses regarding pairs of samples.

Chapter 5

Power of the Tests

5.1 Introduction

In this chapter we present some results concerning the performance of the tests previously described for different alternatives. For the two-sample case, if the alternative against the hypothesis of homogeneity is the shift model $F_1(\mathbf{x}) = F_2(\mathbf{x}+\Delta)$ then the power of a test is the probability to detect the location shift Δ between the two populations. Whaley and Quade (1985) mentioned that this probability depends on five factors:

1. The type I error.
2. The sample sizes.
3. The distribution of each population.
4. The magnitude and the direction of the shift.
5. The test statistic used.

Hotelling's two-sample T^2 test is known to have optimal properties regarding its power against shift alternatives. However, attaining those properties depends on the assumptions of multivariate normality and homogeneity of variance-covariance matrices. Several authors (e.g. Chase and Bulgren (1971), Everitt (1979)) have studied

the robustness of the one and two samples T^2 test with respect to departures from normality. These works point out that for the two samples case, this test is fairly robust in the presence of a non normal distribution. Nevertheless, this is not the case for departures from the assumption of the homogeneity of variance-covariance matrices.

Davis (1980) emphasized the very important point that if the multivariate kurtosis coefficient is greater than the one corresponding to a multivariate normal distribution, (e.g. for a uniform distribution) then the significance levels produced by the Likelihood Ratio Tests (*LRT*) would be substantially reduced. Thus, in this case, the use of this test can be rather dangerous.

Simaika (1941) proved that Hotelling's T^2 two-sample test is uniformly most powerful for multivariate normal observations only when the power depends on the noncentrality parameter alone. This means that if the power of some test depends on more than just the location shift, then the T^2 test is not necessarily more powerful than such a test, even for normal data.

Whaley and Quade (1985), showed that some multivariate runs tests based on graphs constructed by linking points located within a certain distance to each other smaller than a pre-established threshold can have a better performance than the T^2 two-sample test for multivariate normal data. This is due to the fact that the power of the multivariate runs test employed by them does not depend on the noncentrality parameter alone. As the test statistic is based on a threshold distance, the power of the test may vary according to the covariance structures of the samples and to the direction of the shift. For instance, assuming that the shift is positive and that the variables are positively correlated, if the shift is of the *same direction* form: $(\Delta_1, \Delta_2, \ldots, \Delta_p)$ or $(-\Delta_1, -\Delta_2, \ldots, -\Delta_p)$, then we would expect the power of the runs tests to be inferior to the one corresponding to an *opposite direction* shift, i.e. one of the form $(\Delta_1, -\Delta_2, \Delta_3, \ldots, \Delta_{p-1}, -\Delta_p)$.

Whaley and Quade (1985) also pointed out that, for the two-sample case, the power of multivariate runs tests is likely to be increased if graphs like the *n-MST* were used instead of those generated by threshold distances. For this reason, and also in view of the degree of arbitrariness implied by choosing a threshold, we did not use graphs

based on that approach.

As a first example, we now discuss an experiment similar to those published by Friedman and Rafsky (1979) and Schilling (1986). We used these examples in order to "calibrate" our programs by contrasting our results with published ones. We generated 100 replications of two samples of standard multivariate normal random variables. The dimensions considered were $p = 1, 2, 5, 10, 20$; the sample sizes used in this example were $n_1 = n_2 = 100$, so the normal approximation to the null distribution of Γ_R should be expected to work very well.

In the first part of this example, we examine the performance of Γ_R using several graphs based on Euclidean distances as $\mathcal{G}_X$ for testing the hypothesis of homogeneity against location alternatives.

Table 5.1 shows the power of the runs test based on Γ_R for several graphs. The separations, in Euclidean distance, between the mean vectors of the two samples for several dimensions are equal to those used by Friedman and Rafsky (1979) as well as by Schilling (1986); we worked with them in order to have a point of comparison with the only published results obtained for some graphs. They are (0.3, 0.5, 0.75, 1.0, 1.2) corresponding to p=(1, 2, 5, 10, 20). Throughout this chapter we used 100 simulations to estimate the power of the tests, so the conclusions should be taken with some caution. The significance level used in all the examples was 5%; the tables show the number of times which the test statistics exceeded the corresponding critical value for that significance level.

The *LRT* results reported in Table 5.1 were obtained with the Bartlett-Nanda-Pillai trace criterion; we used the approximation suggested in Mijares (1990). For this sample size, the use of other parametric test statistics (e.g. Wilk's Λ) makes no substantial difference on the estimated power.

Tests	p				
	1	2	5	10	20
LRT	45	77	91	92	84
1-*MST*	13	19	36	64	70
2-*MST*	14	30	52	73	98
3-*MST*	15	33	66	85	90
7-*MST*	22	43	79	96	99
8-*MST*	19	46	85	97	99
1-*NNG*	12	18	37	57	66
$T^*_{1N}(C)$	5	7	41	47	
$V^*_{1N\mathrm{W}}(C)$	71	100	99	100	
2-*NNG*	12	27	51	72	88
$T^*_{2N}(C)$	9	22	57	67	
$V^*_{2N\mathrm{W}}(C)$	68	100	99	100	
3-*NNG*	13	32	65	87	89
$T^*_{3N}(C)$	12	27	74	76	
$V^*_{3N\mathrm{W}}(C)$	72	100	97	100	
7-*NNG*	20	41	82	97	97
8-*NNG*	18	42	83	98	99
1-*RNG*	13	25	76	90	100
2-*RNG*	16	49	97	91	98
1-*GG*	13	25	90	100	100
2-*GG*	16	49	97	98	100

Table 5.1: Power against location alternatives for two samples

The rows marked as T^*_{nN} and $V^*_{nN\mathrm{W}}$ are Schilling (1986) two-sample n- nearest neighbour statistics, defined as:

$$T^*_{nN} = \frac{1}{n \cdot N} \sum_{i=1}^{N} \sum_{r=1}^{n} I_i(r)$$

and

$$V^*_{nN\mathrm{W}} = \frac{1}{n \cdot N} \sum_{\alpha=1}^{2} \sum_{i \in \Omega_\alpha} w_\alpha(Z_i) \sum_{r=1}^{n} I_i(r),$$

where Ω_α represents the α-th sample and $I_i(r)$ is an indicator random variable which takes the values 1 if the point i and its r-th nearest neighbour share the same sample label and 0 otherwise. $w_\alpha(Z_i)$ are the optimal weights defined by Schilling (1986). The first statistic corresponds to the unweighted proportion of all pairs of n nearest neighbours which come from the same sample. The second is one of its weighted versions proposed by Schilling (1986). Although Schilling's tests seem very powerful indeed, it is not clear how to generalize the optimal weights defined by this author to the K-sample case. Also, this test might prove to be not as powerful against scale alternatives.

The results presented in Table 5.1 closely resemble those obtained by Friedman and Rafsky (1979). There is a marked effect of the number of orthogonal graphs on the power of the runs tests. Also, it can be seen that the tests based on the *RNG* s and the *GG* s have a more parsimonious performance than those based on *MST* s and *NNG* s, in the sense that they attain comparable power levels using less edges. Another remark concerns the dimensionality of the data: the power of the runs tests seems to improve for higher dimensions; this is contrary to the performance of the parametric tests (Kshirsagar (1972)).

So far we have only given an example of the comparisons we are interested in. In the next sections we cover a wider range of tests and multisample multivariate situations. We must rely on simulation studies which will be very limited, due to the many factors that affect the behaviour of the nonparametric multivariate tests.

5.2 Shift Alternatives

Our first example involves 8 samples, each of size 25. The shift Δ was applied to only one variable in one sample; the rest of the samples had the same mean. All the samples were generated with the $p \times p$ identity as their variance-covariance matrix. We must stress that we did not consider the power of the likelihood ratio tests as a benchmark to make the comparisons between the parametric and nonparametric tests. In order to do so, we would have had to obtain the power of the nonparametric tests using the Δ

value which prescribes a determinate power level for a *LRT*. However, as the power of the multivariate runs and ranks tests does vary in accordance with the direction of the location shift, this would have lead to consider many complicated patterns in the generation of the shifts. Thus, we worked in a situation which might be thought to be somehow disadvantageous for the nonparametric tests.

The *LRT* column presents the results for the *LRT* criterion which achieved the largest power. The criteria used were the Bartlett-Nanda-Pillai trace criterion using the approximation suggested by Mijares (1990) and Wilk's Λ, using Bartlett's χ^2 and Box's F approximations as described by Anderson (1984). For normal data, all these approximations are reported to be remarkably accurate, even for small sample sizes.

The column marked as *P-S* holds the results for the three Puri-Sen multivariate rank tests that we considered for location alternatives. They are *MMRST*, *MMMT*, and *MMNST*, and were described in Section 3.4. Similarly to the *LRT* case, we chose the test that yielded the largest power amongst these three tests.

The three columns under *Runs* show the power estimates obtained for 1-*MST*, 10-*MST* and 15-*MST*. The corresponding *NNG*s produced very similar results. The largest difference in power was usually observed between the 1-*MST* and the 5-*MST*. Using 1-*RNG*s or 1-*GG*s produced results comparable to those obtained using between 5 and 10 orthogonal *MST*s.

Using higher order *RNG*s or *GG*s does increase the power. However, it is always possible to obtain similar or better results by adding orthogonal *MST*s, and so we do not report the results obtained with those graphs. Clearly, we could go on until we had a complete or almost complete graph. This would happen particularly for *n-RNG*s or *n-GG*s for points in higher dimensions or for a relatively large number of orthogonal *MST*s or *NNG*s for small samples. Considering the complete graph as a basis for the runs statistics does not necessarily lead to an improvement in their power. This is because the runs test rejects the hypothesis of homogeneity whenever the number of links defined by points from different samples is significantly small. For instance, consider two samples, each of size 4 differing only in location. There are 28 links in the complete graph -which is equivalent to 4-*MST*. In this graph there would be 16 links

from different samples against 12 from the equal sample, and thus the null hypothesis would never be rejected.

The last column of this table refers to the estimated power of the nonparametric rank tests based on the Friedman-Rafsky multivariate ranking procedures. Again, we present the largest power attained amongst all the tests. The performance of these tests is very poor and, in general, it decreases as the number of dimensions grows.

The power levels attained with the tests based on degree 1 nodes were so low in this example that we do not report them.

From Table 5.2 we can see that the Puri-Sen tests are a safe option, attaining power levels similar or even higher than those corresponding to the *LRT* for almost all the cases reviewed. For normal data it is clear that the *LRT* is the best procedure to use, as we would expect. The high power levels achieved by the *LRT* for the uniform distribution should be considered bearing in mind Davis' remarks. For the lognormal distribution, the runs tests with a large enough number of orthogonal *MST*s have the best performance.

We cannot affirm whether or not the multisample nonparametric tests increase their power levels with the dimensionality of the observations, as Friedman and Rafsky (1979) noted for their two-sample tests. It is obvious that the number of samples has an effect on their performance, but it does not seem easy to describe it.

We now turn our attention to smaller sample sizes. The tests based on the Friedman-Rafsky multivariate ranks performed very poorly indeed in these examples and so we do not include them in the following tables.

Table 5.3 has a similar pattern to that of Table 5.2. Again, the *LRT* is the best option for normal data, the Puri-Sen tests produce reasonably high power levels and the runs tests perform much better only for lognormal data,

In Table 5.4 we show an example with N=50 and four samples. We present the results obtained for 10-dimensional data, which exhibit the same behaviour observed before. We would regard suspiciously the high power levels attained by the *LRT* for uniform data. It seems that the Puri-Sen tests' performance deteriorates for higher dimensions, while the runs tests steadily achieve good levels for large numbers of orthogonal graphs.

Normal data; N=200, K=8, n_i=(25, 25, 25, 25, 25, 25, 25, 25)

p	Δ	*LRT*	*P-S*	*Runs*			*F-R*	p	Δ	*LRT*	*P-S*	*Runs*			*F-R*
2	0.4	21	18	8	12	16	12	5	0.4	14	10	5	5	9	12
	1.2	96	95	29	59	72	21		1.2	87	82	20	35	41	18
10	0.4	10	9	5	3	7	17	20	0.4	10	12	6	5	2	17
	1.2	64	59	19	27	32	11		1.2	49	43	14	23	33	13

Uniform data; N=200, K=8, n_i=(25, 25, 25, 25, 25, 25, 25, 25)

p	Δ	*LRT*	*P-S*	*Runs*			*F-R*	p	Δ	*LRT*	*P-S*	*Runs*			*F-R*
2	0.2	58	67	14	22	22	21	5	0.2	33	36	9	9	15	6
	0.4	97	95	17	43	50	16		0.4	100	100	45	56	62	25
10	0.2	23	27	4	8	14	8	20	0.2	9	10	6	8	12	6
	0.4	85	79	20	32	38	14		0.6	95	88	28	68	93	16

Exponential data; N=200, K=8, n_i=(25, 25, 25, 25, 25, 25, 25, 25)

p	Δ	*LRT*	*P-S*	*Runs*			*F-R*	p	Δ	*LRT*	*P-S*	*Runs*			*F-R*
2	0.4	23	34	9	14	19	15	5	0.4	14	18	9	17	15	9
	1.2	99	100	40	94	99	45		1.2	85	100	25	70	89	21
10	0.4	8	15	5	9	9	12	20	0.4	7	10	5	9	8	9
	1.2	60	80	19	53	57	20		1.2	43	54	17	31	34	11

Lognormal data; N=200, K=8, n_i=(25, 25, 25, 25, 25, 25, 25, 25)

p	Δ	*LRT*	*P-S*	*Runs*			*F-R*	p	Δ	*LRT*	*P-S*	*Runs*			*F-R*
2	0.4	28	36	9	17	25	20	5	0.4	11	19	8	16	14	14
	1.2	99	100	51	94	99	71		1.2	91	99	54	98	99	54
10	0.4	15	20	9	14	11	9	20	0.4	11	14	10	12	8	7
	1.2	86	94	62	90	95	29		1.2	58	80	40	87	93	20

Table 5.2: Power against location alternatives for eight samples

Normal data; N=30, K=4, n_i=(7, 7, 8, 8)

p	Δ	*LRT*	*P-S*	*Runs*			p	Δ	*LRT*	*P-S*	*Runs*		
2	0.9	22	22	9	23	33	5	0.9	21	13	14	15	26
	2.0	87	74	38	77	82		2.0	73	40	36	68	69

Uniform data; ; N=30, K=4, n_i=(7, 7, 8, 8)

p	Δ	*LRT*	*P-S*	*Runs*			p	Δ	*LRT*	*P-S*	*Runs*		
2	0.4	59	43	21	40	46	5	0.4	34	19	13	34	40
	0.9	100	100	64	100	100		0.9	97	73	60	98	99

Exponential data; shift alternative; N=30, K=4, n_i=(7, 7, 8, 8)

p	Δ	*LRT*	*P-S*	*Runs*			p	Δ	*LRT*	*P-S*	*Runs*		
2	0.4	16	18	12	14	25	5	0.4	11	10	6	15	16
	2.0	90	79	52	98	99		2.0	71	69	42	68	77

Lognormal data; shift alternative; N=30, K=4, n_i=(7, 7, 8, 8)

p	Δ	*LRT*	*P-S*	*Runs*			p	Δ	*LRT*	*P-S*	*Runs*		
2	1.3	60	81	42	92	90	5	1.3	34	31	39	77	71
	2.0	81	98	59	99	100		2.0	45	67	58	100	100

Table 5.3: Power against location alternatives for four samples

N=50, K=4, n_i=(13, 13, 12, 12)

distribution	p	Δ	*LRT*	*P-S*	*Runs*		
Normal	10	0.9	16	10	5	13	21
		1.7	69	43	27	66	69
Uniform	10	0.1	12	10	6	9	11
		0.4	58	23	14	29	42
Exponential	10	0.9	21	37	12	15	16
		1.7	63	69	34	76	73
Lognormal	10	0.9	20	26	16	49	58
		1.3	29	48	46	90	97

Table 5.4: Power against location alternatives for four samples

As we said in the introduction to this chapter, the power of the nonparametric tests that we are considering depend on a collection of causes. In our last example for location alternatives, we present an example of the impact of some of these factors on the performance of the tests. The results appear in Table 5.5

We divide 100 data sets of size 30 into 5 subsamples. For $p = 2$ we consider two possible partitions: (6, 6, 6, 6, 6) and (6, 5, 7, 5, 7); they appear in the table under "eq" and "uneq" sample sizes, respectively. We consider two possibilities with respect to the variances of the samples; they are (1, 1, 1, 1, 1) and (1, 2, 3, 2, 1) and are indicated in the table as "eq" and "uneq", respectively.

We assume that Δ=0.6 for $p = 2$ and $p = 5$.

Normal data; p=2, N=30, K=5

sizes	σ^2	*LRT*	*P-S*	1-*MST*	10-*MST*	15-*MST*	*RNG*	*GG*	*F-R*
eq	eq	20	16	9	19	20	9	8	6
eq	uneq	14	15	11	17	10	23	24	11
uneq	eq	23	12	8	22	30	13	15	12
uneq	uneq	15	16	21	27	18	22	38	26

Normal data; p=5, N=30, K=5

sizes	σ^2	*LRT*	*P-S*	1-*MST*	10-*MST*	15-*MST*	*RNG*	*GG*	*F-R*
eq	eq	23	13	12	25	29	15	15	4
eq	uneq	8	6	31	2	3	39	61	52

Table 5.5: Power against location alternatives for five samples

No clear pattern emerges from this table for the *LRT* or the Puri-Sen tests. However, we may notice that, in general, the nonparametric tests perform better than those tests in the absence of equal variances. Also, there are some results which show that one should be careful with the number of orthogonal graphs used to construct the runs test statistic. If this number is too large with respect to the sample sizes, then too many edges from the same sample would be included and as a result the hypothesis of homogeneity will be incorrectly accepted. To correct this we suggest to use the *RNG* or *GG*, as shown here. Although these graphs produce test statistics which seem to have slightly lower power than those obtained with a large numbers of *NNG*s or *MST*s, they comprise a more

robust possibility against picking up too many redundant edges. Of course, for higher dimensions, the *GG* might be very similar to the complete graph. Thus, we recommend the *RNG* as a good compromise between power and generality of application.
We can also note that the multivariate rank tests (in particular Kiefer's one) seem to have good power. However, as we are using the approximation discussed in Subsection 3.3.4 assuming that it would be good enough for the present sample structure, these results should be taken with due caution.

5.3 Scale Alternatives

We now present a few experiments for tests concerning the variance-covariance matrix. For our first example, we change the scale of only one sample and use normal data in 2, 5, 10 and 20 dimensions. The total number of observations is 200, partitioned in 8 samples, each of size 25. The results appear in Table 5.6.

Normal data
N=200, K=8, n_i=(25, 25, 25, 25, 25, 25, 25, 25)

p	σ^2	*LRT*	*P-S*	deg 1	*Runs*	F-R
2	1.2	51	52	9	20	29
	1.4	99	100	14	72	83
5	1.2	66	84	58	41	44
	1.4	100	100	90	89	88
10	1.1	14	42	36	14	20
	1.2	74	99	82	59	51
20	1.08	13	53	66	18	22
	1.15	37	97	94	54	51

Table 5.6: Power against scale alternatives for eight samples

The *LRT* column corresponds to Box's test for homogeneity of variance-covariance matrices. We used the F approximation described by Anderson (1984). The *P-S* column holds the results obtained with the $(MMNST)^2$. We include the power levels concerning the degree 1 test. As we remarked in Chapter 4, this test seems to be an adequate choice to test scale differences.

The results presented in the *Runs* column were obtained using 5-*MST*. The last column refers to the rank test which achieved the largest power level in each particular case with the Friedman-Rafsky radial ranking. We should remark that Kiefer's test performed slightly better than the rest of the tests.

It is clear that Box's is not the best choice for higher dimensional data. The Puri-Sen test has a steady response in all dimensions, even for small scale differences. It is also worth noting that the performance of the multivariate ranks tests based on the 1-*MST* is worse for higher dimensions, as the representation of the distance structure obtained with the radial ranking becomes less accurate.

The runs tests are geared more towards detecting shift differences. Nevertheless, they performed well enough for $p \leq 10$ in this case.

For our next example, we considered $N = 30$, and $K = 4$. The sample sizes were 7, 7, 8, and 8; we generated points in 2, 5, and 10 dimensions. The results appear in Table 5.7. The *Runs* columns correspond to 1, 10 and 15 *MST*s. The latter graph is, in this case, the complete graph. The very low power levels obtained with it illustrate a point which we have mentioned before: that merely increasing the number of edges does not necessarily lead to a more effective test statistic.

Normal data; scale alternative; $N = 30$; n_i=(7, 7, 8, 8); K=4

p	σ	*Box*	*P-S*	deg 1	*Runs*			*F-R*
2	2.2	64	52	11	7	25	8	29
	3.4	92	84	9	17	44	6	34
5	2.2	62	55	16	14	34	7	27
	3.4	97	85	21	15	47	4	50
10	1.8			23	10	25	3	25
	2.2			31	15	28	5	35

Table 5.7: Power against scale alternative for four samples

In actual fact, this situation was observed in several cases, even when the *n-MST* considered did not coincide with the complete graph. For $p = 10$, we see that the degree-1 and the *F-R* radial ranking tests attain better power levels than the runs tests. The effect of using the complete graph in the latter tests is patent.

We performed other experiments to study scale alternatives with other distributions, but the the best results, were obtained either with the Box test or with the $(MMNST)^2$. As the latter seems to be a very powerful and robust test, we would recommend using another nonparametric test for scale alternatives only when the sample structure is such that does not allow to use this Puri-Sen type test with confidence. In those cases we would have to insure that the distribution used is adequate. In general, using four-moments approximations or sampling from the exact permutational distribution should guarantee that.

Chapter 6

Association and Prediction Measures

6.1 Introduction

Friedman and Rafsky (1983) outlined another application of graph theoretic concepts in multivariate data analysis. They proposed a measure of association and one of prediction taking advantage of the theory of generalized correlation coefficients (*GCC*); some aspects of this theoretical framework have been already discussed in Section 2.3. In this chapter we explore the possibilities of such measures.

Let us consider a sample of size N of ordered pairs $(\mathbf{x}_i, \mathbf{y}_i)$ from (possibly multivariate) random variables **X** and **Y**. If a_{ij} denotes a score measured for the i-th and j-th observations over the **X**-values, and b_{ij} is another score for these observations over the random variable **Y**, then a statistic of the form $\Gamma = \sum_i^N \sum_j^N a_{ij}\, b_{ij}$ is a *GCC*. The exact and asymptotic distributions for this class of statistics have been discussed in Section 2.3. It was shown there that, under the null hypothesis of no correlation for **X** and **Y**, Γ is asymptotically normally distributed for a wide variety of score functions a_{ij} and b_{ij}. We considered in detail a particular case of Γ: the one which arises by considering one of the scores to be directly related to sample identity and the other to be based on the edges of an interpoint distance graph. We studied this situation in the context of testing if K multivariate samples had the same underlying distribution. If we can reject the hypothesis of no correlation for these score functions, then we conclude that the samples are not homogeneous.

We now turn our attention to the general problem of determining to what extent points which have similar values in the **X** space correspond to observations which are near in the **Y** space, based on a sample of N ordered pairs $(\mathbf{x}_i, \mathbf{y}_i)$. We follow the two approaches presented by Friedman and Rafsky (1983). In the next two sections, we describe their multivariate measures of association and prediction.

6.2 A Measure of Association

Let $\mathcal{G}_X$ and $\mathcal{G}_Y$ be any two spanning graphs constructed over **X** and **Y**. If a_{ij} and b_{ij} are indicator variables taking the value of unity if the i-th and j-th points form an edge in $\mathcal{G}_X$ and $\mathcal{G}_Y$, respectively, then

$$\Gamma_1 = \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} a_{ij}\, b_{ij} \tag{6.1}$$

is the number of edges in the intersection of both graphs. If $\mathcal{G}_X$ and $\mathcal{G}_Y$ are graphs like those discussed in Chapter 1, i.e. graphs whose edges correspond to pairs of points which are somehow near, then we should reject the hypothesis of no correlation between **X** and **Y** whenever the observed value of Γ_1 is too large or too small; the former case indicates the presence of a strong positive relation between **X** and **Y** and the latter, a negative one. From the previous chapters we have that for the K-sample case it only makes sense to reject the null hypothesis if the value of Γ_1 is too small, implying that the number of pairs of points having the same sample identities and forming an edge in the interpoint distance graph is significantly smaller than it would be expected if the sample identities would be assigned by random labelling.

The null distribution for Γ_1 can be approximated with a sample from the permutational distribution or with a Normal approximation. We did not attempt to construct approximations based on the first four moments. The work presented in Chapter 2 for the K-sample multivariate runs statistic was considerably eased by the fact that $\mathcal{G}_Y$ was considered to be $\cup_{j=1}^{K} \mathcal{K}_{n_j}$; we do not have this advantage for the general case.

Let z_i be and indicator variable taking the value of 1 if the i-th edge of $\mathcal{G}_X$ is also an edge of $\mathcal{G}_Y$.

Therefore $\Gamma_1 = \sum_{i=1}^{e_X} z_i$, and the first moment of Γ_1 is:

$$E(\Gamma_1) = \sum_{i \in \mathcal{G}_X} \Pr[z_i = 1] = e_X p = \frac{e_X e_Y}{\binom{N}{2}} \tag{6.2}$$

where $p = \Pr[z_i = 1] = e_Y \binom{N}{2}^{-1}$. From equation (6.2) we have that the expected value of the number of edges in the intersection of both graphs is the number of edges in $\mathcal{G}_X$ times the probability that any edge of $\mathcal{G}_X$ appears also in $\mathcal{G}_Y$.

Then $\text{var}(\Gamma_1)$ is calculated exactly as in equation (2.28) in Chapter 2:

$$\text{var}(\Gamma_1) = \text{var}\left(\sum_{i=1}^{e_X} z_i\right) = \sum_{i=1}^{e_X} \text{var}(z_i) + 2 \sum_{i<j} \text{cov}(z_i, z_j)$$

so we can write

$$\begin{aligned} \text{var} \quad (\Gamma_1 \mid e_X, e_Y, C_X, C_Y) \; &= \frac{2 e_X e_Y}{N(N-1)} \left\{ 1 - \frac{2 e_X e_Y}{N(N-1)} \right\} \\ &+ \frac{4}{N(N-1)(N-2)} \\ &\cdot \left[C_X C_Y + \frac{\{e_X(e_X-1) - 2C_X\}\{e_Y(e_Y-1) - 2C_Y\}}{N-3} \right] \end{aligned} \tag{6.3}$$

Lefkovitch (1984, 1985) worked with a statistic based on the intersection of two *RNG*s. He discussed several interesting analysis for problems involving establishing association between two distance matrices. For instance, he was concerned with assessing the correlation between the geographical distribution and some attributes measured over plants in a region. This author follows a 2×2 contingency table approach similar to that developed by Barton and Davis (1966). Given two dissimilarity matrices constructed for N observations, $\mathbf{D}_v$ and $\mathbf{D}_w$, and their *RNG*s, let $|\mathbf{E}_v|$ and $|\mathbf{E}_w|$ denote the cardinality of their respective edge sets. Then, Lefkovitch forms the following contingency table

$\lvert \mathbf{E}_v \cap \mathbf{E}_w \rvert$	$\lvert \mathbf{E}_v \setminus \mathbf{E}_w \rvert$
$\lvert \mathbf{E}_w \setminus \mathbf{E}_v \rvert$	$\lvert \mathbf{K} \setminus (\mathbf{E}_v \cap \mathbf{E}_w) \rvert$

where $|\mathbf{K}|$ denotes the cardinality of the complete graph with N points. If there is no

association between $\mathbf{D}_v$ and $\mathbf{D}_w$, then the relative neighbours in of each of the objects in $\mathbf{E}_v$ would tend to be independent of those in $\mathbf{E}_w$. On the other hand, if $|\mathbf{E}_v \cap \mathbf{E}_w|$ is greater than its expectation under the null hypothesis, then there would be evidence that the relative neighbours tend to be alike, which would indicate that $\mathbf{D}_v$ and $\mathbf{D}_w$ have some common information. This hypothesis may be tested using an ordinary log-likelihood test for marginal independence, whose test statistic, G^2 has, asymptotically, a χ^2 distribution with 1 df. We do not pursue this approach here.

Friedman and Rafsky (1983) mentioned that it is the lack of regard for large distances in the computation of Γ_1 that gives rise to its good power characteristics against general alternatives.

Although Γ_1 is a useful statistic which should have good performance against a great variety of alternatives, these authors mentioned that if the relationship between **X** and **Y** is not one-one, this generality could mean some loss of power. If a many-one relationship constitutes the alternative hypothesis, then it is important to measure to what extent the observed values of **X** can be used to predict values of **Y**, without any consideration about how well could values of **X** be predicted from **Y**.

A more powerful test for those situations should make use of the hypothesized relationship between **X** and **Y**, and thus would involve only small interpoint distances from **X** while including both small and large distances from **Y**. We discuss such a test in the following section.

6.3 A Measure of Prediction

Let us assume that $\mathcal{G}_X$ links points which are somewhow close to each other. We now define the score $R_i(j)$ to be the position of the j-th observation in the list resulting from increasingly ordering the sample points according to their **Y**-distances from the i-th observation. We take $R_i(i) = 0$ and so $1 \leq R_i(j) \leq N-1$ for all i and j.

Considering $a_{ij} = 1$ if $(i,j) \in \mathcal{G}_X$ and $a_{ij} = 0$ otherwise, and $b_{ij} = R_i(j)$, the Friedman-

Rafsky measure of prediction can be written as

$$\Gamma_2 = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}\ b_{ij} \quad = \quad \sum_{(i,j)\in \mathcal{G}_X} R_i(j). \tag{6.4}$$

We have that if two nodes, i and j, define an edge in $\mathcal{G}_X$ the corresponding $R_i(j)$ should take a small value and then, rejection of the null hypothesis of no correlation is indicated by small values of Γ_2.

If $b_{ij} = R_i(j)$, then the degrees of the spanning graph associated with these scores are $d_i = N(N-1)$, for all the nodes. This assures that Daniels' condition is satisfied, as

$$\lim_{N\to\infty} \frac{\left(\sum_{i=1}^{N} d_i^3\right)^2}{\left(\sum_{i=1}^{N} d_i^2\right)^3} = 0,$$

the numerator being $\mathcal{O}(N^{14})$ and the denominator $\mathcal{O}(N^{15})$.

This implies that the asymptotic normality of Γ_2 depends entirely on $\mathcal{G}_X$ satisfying the conditions established in equation (2.18).

Friedman and Rafsky (1983) obtained the first two moments of the null distribution of Γ_2. We now present their formulae.

From equation (6.4),

$$\mathrm{E}(\Gamma_2) = \sum_{(i,j)\in \mathcal{G}_X} \mathrm{E}\,[R_i(j)] = 2\,e_X\,\mathrm{E}\,[R_i(j)] \tag{6.5}$$

$$\mathrm{E}\,[R_i(j)] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_i(j) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N-1} j = \frac{N}{2} \tag{6.6}$$

From these two equations, we have that

$$\mathrm{E}(\Gamma_2) = N\,e_X \tag{6.7}$$

Obtaining an expression for the second moment is more complicated. We first have

that

$$\text{var}(\Gamma_2) = \sum_{(i,j) \in \mathcal{G}_X} \text{var}\,[R_i(j)] + \sum_{\substack{(i,j),\,(k,l) \in \mathcal{G}_X \\ (i,j) \neq (k,l)}} \text{cov}\,[R_i(j),\, R_k(l)] \tag{6.8}$$

and

$$\begin{aligned} \text{var}\,[R_i(j)] &= \text{E}\left[R_i(j)^2\right] - \{\text{E}\,[R_i(j)]\}^2 \\ &= \frac{1}{N-1}\sum_{j=1}^{N-1} j^2 - \frac{N^2}{4} \\ &= \frac{N(2N-1)}{6} - \frac{N^2}{4} = \frac{N(N-2)}{12} \end{aligned} \tag{6.9}$$

The second sum in equation (6.8) can be written as

$$\text{cov}\,[R_i(j),\, R_k(l)] = \text{E}\,[R_i(j)\; R_k(l)] - \frac{N^2}{4} \tag{6.10}$$

In order to calculate the expectation of $[R_i(j)\; R_k(l)]$, we have to consider 6 cases of ordered edge pairs defined by nodes (i, j) and (k, l); these are:

Case 1:	$R_i(j)$	$R_i(k)$	$i \neq j \neq k$
Case 2:	$R_i(j)$	$R_k(j)$	$i \neq j \neq k$
Case 3:	$R_i(j)$	$R_j(k)$	$i \neq j \neq k$
Case 4:	$R_i(j)$	$R_k(i)$	$i \neq j \neq k$
Case 5:	$R_i(j)$	$R_j(i)$	$i \neq j \neq k$
Case 6:	$R_i(j)$	$R_k(l)$	$i \neq j \neq l$

Friedman and Rafsky defined the following two parameters of the matrix $\mathbf{R} = [R_i(j)]$ which are needed to calculate the expressions involved in some of these cases:

$$A_R = \sum_{i=1}^{N}\sum_{j=1}^{N} R_i(j)\; R_j(i) \qquad \text{and} \qquad B_R = \sum_{i=1}^{N}\left(\sum_{j=1}^{N} R_i(j)\right)^2 \tag{6.11}$$

The expressions required for each case appear in an appendix to the paper by Friedman

and Rafsky (1983). The expression for the variance of Γ_2 may be written as:

$$\text{var} \quad (\Gamma_2 | e_X, C_X, A_R, B_R) = \frac{e_X^2}{N-3}\left\{\frac{N(3N+1)}{3} + \frac{4(A_R - B_R)}{N(N-1)(N-2)}\right\}$$

$$+ \frac{e_X}{N-3}\left\{\frac{2(N-1)(N-4)A_R + 4B_R}{N(N-1)(N-2)} - \frac{N,(N-1)(N-2)}{3}\right\}$$

$$+ \frac{2C_X}{(N-2)(N-3)} \cdot \qquad (6.12)$$

$$\left\{\frac{(N+1)B_R - 2(N-1)A_R}{N(N-1)} - \frac{N,(3N-4)(N^2-1)}{12}\right\}$$

The extent to which a normal approximation based on the first two moments follows the null distribution of Γ_2 was not discussed by Friedman and Rafsky (1983). However, they gave an example in which that approximation seems to work adequately for N=100. For small sample sizes, one could follow the two approaches outlined in Chapter 2. The easiest alternative seems to be sampling from the permutational distribution of Γ_2 conditional on $\mathcal{G}_X$. It is possible to construct the expressions for the third and fourth moments of this distribution. However, this approach seems rather cumbersome, as the number of cases that is necessary to consider increases considerably for the higher moments, so we did n t pursue it. In the next section we explore the adequacy of using a Normal approximation for Γ_1 and Γ_2.

6.4 Approximations to the Null Distributions of Γ_1 and Γ_2

We are interested in assessing the performance of the Normal approximations described in the previous sections for relatively small sample sizes and for a variety of graphs. To do so, we compare this approximations with results produced by sampling from the exact permutational distribution. The mean and the variance of the normal approximations are calculated with expressions (6.2) and (6.3) for Γ_1 and (6.7) and (6.13) for Γ_2, respectively.

We start with some examples for Γ_1. We constructed two independent samples of

multivariate normal values and calculated their Euclidean distance matrices. Table 6.1 presents the results for several combinations of graphs, sample size and dimension. The columns μ and σ correspond to the mean and standard deviation of the normal approximation, while $\bar{x}$, s, b_1, b_2 and κ are the mean, standard deviation, the coefficients of skewness and kurtosis and the Pearson criterion estimated from samples of size 1000 from the exact permutational distribution. Figure 6-1 shows an example of an apparently nonnormal exact distribution with the density estimate obtained with 1000 realizations from the exact permutational distribution of Γ_1 and the corresponding normal approximation. The density estimate was obtained with a Gaussian kernel using the optimal smoothing parameter proposed by Silverman (1986, §3.4). The normal approximation is shown in the solid line. From Table 6.1 and Figure 6-1 we can see

Graphs	N	p	μ	σ	$\bar{x}$	s	b_1	b_2	κ
1-*MST*- 1-*MST*	50	2	1.960	1.359	1.964	0.966	0.279	3.483	1.763
1-*MST*- 1-*MST*	50	20	1.960	1.344	1.909	0.936	0.192	3.061	-0.330
1-*MST*- 1-*MST*	200	2	1.990	1.400	1.943	0.998	0.266	3.304	-1.131
1-*MST*- 1-*MST*	200	20	1.990	1.396	1.963	1.002	0.398	3.705	1.518
1-*MST*- 1-*MST*	500	2	1.996	1.409	1.996	1.022	1.263	5.521	0.988
1-*MST*- 1-*MST*	500	20	1.996	1.407	1.9876	0.9683	0.019	2.253	-0.009
1-*MST*- 1-*GG*	50	2	3.640	1.817	3.526	1.242	0.131	3.208	4.866
1-*MST*- 1-*GG*	50	20	38.400	2.817	39.030	1.895	-0.017	2.854	0.005

Table 6.1: Γ_1 approximations: parameter values

that if one of the graphs considered is a sparse one then the normal approximation does not fit well the null permutational distribution of Γ_1 even for sample sizes as large as 500. We obtained better results considering denser graphs. For instance, the *GG*s used with N=50 have 91 and 960 edges for 2 and 20 dimensions, respectively. The 1-*MST* has always 49 edges. Thus, the corresponding proportions of edges from the complete graph are 0.04, 0.0743 and 0.78 for the 1-*MST* and the two *GG*s, respectively. The *RNG* or a moderate number of orthogonal *MST*s or *NNG*s are sensible choices to calculate Γ_1, as has been suggested by Lefkovitch (1984).

Table 6.1 shows that the variance calculated with expression (6.3) is always larger than the one estimated from the permutational null distribution. Thus, the normal

approximation suggested by Friedman and Rafsky (1983) would yield a conservative test. However, we would recommend using a large enough sample from the exact distribution in order to obtain better approximations to the true significance level. It

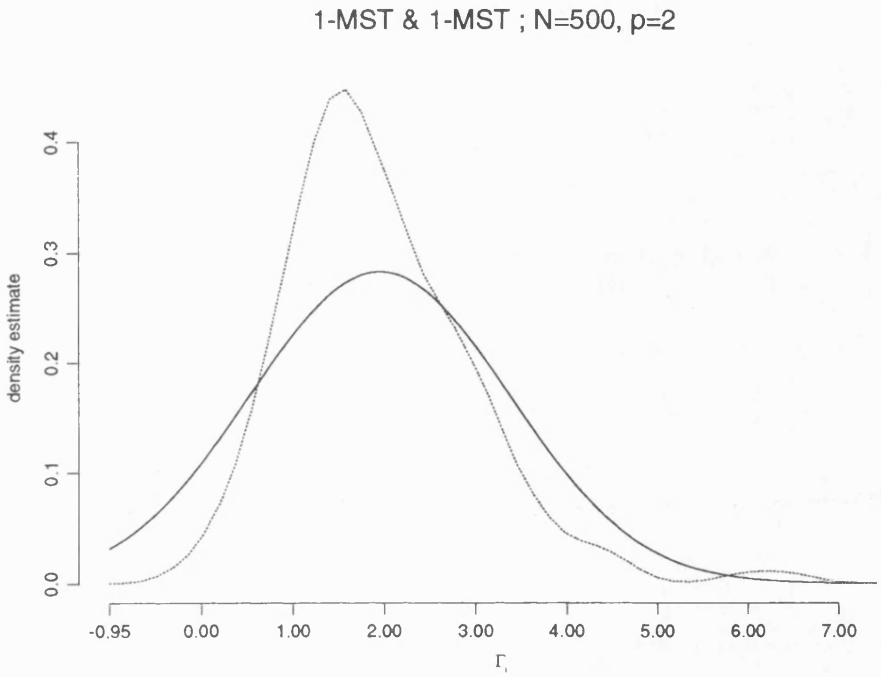


Figure 6-1: Γ_1 Approximations

is not surprising that the normal approximation works better for Γ_2 than for Γ_1, as the former statistic has a wider range.

For Γ_2, we have that even for sparse graphs like 1-*MST* and sample sizes as small as 20 we always observe that the exact permutational distribution closely resembles a normal distribution. Table 6.2 and Figures 6-2 and 6-3 show some examples of this fact. We find that the variance of Γ_2 calculated using equation (6.8) is always larger than the variance estimated from the samples from the permutational distribution. This leads, as for Γ_1, to a conservative test if we apply the normal approximation directly. Friedman and Rafsky (1983) pointed out that Γ_1 considers only some selected, relatively small distances without regard for the larger distances, while Γ_2 makes a fuller use of the distance matrix. So, as tests based on Γ_1 are appropiate in a wide variety of cases, the price paid for this generality is less power in those situations in which large distances

Graphs	N	p	μ	σ	$\bar{x}$	s	b_1	b_2	κ
1-*MST*	50	20	2450	179.37	2449.6	137.92	0.013	3.201	0.021
5-*MST*	50	20	12250	390.65	12259	270.71	-0.002	2.987	0.067
1-*MST*	20	20	380	37.75	380	31.93	-0.001	2.774	0.003
1-*GG*	20	5	1580	64.61	1580.5	50.98	-0.000	2.966	0.001

Table 6.2: Γ_2 Approximations: Parameter Values

play a significant role in defining the common information shared by both distance matrices. We think that using one or two sufficiently dense graphs in Γ_1 would provide an adequate compromise between generality and power.

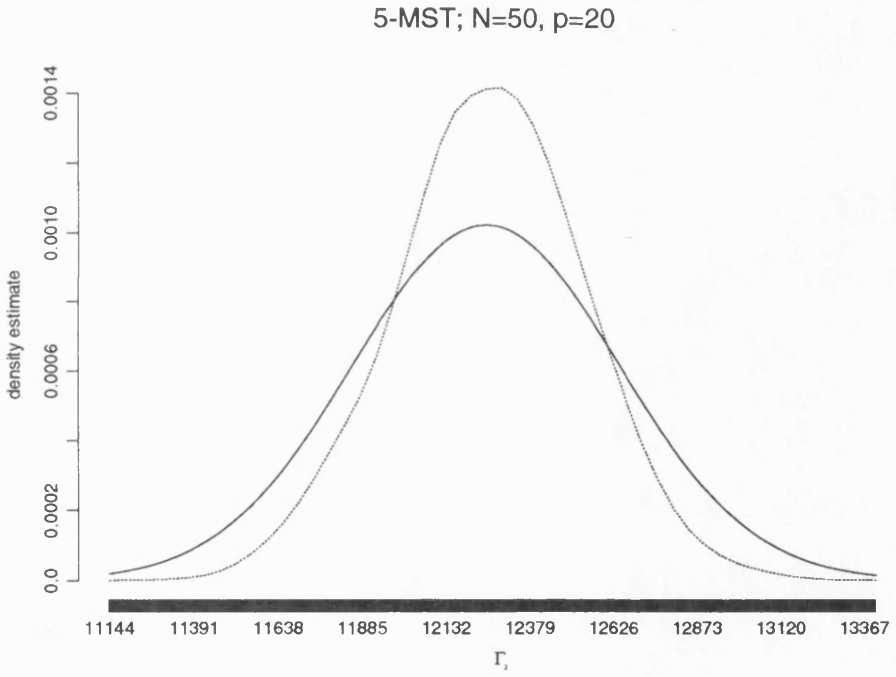


Figure 6-2: Γ_2 Approximations

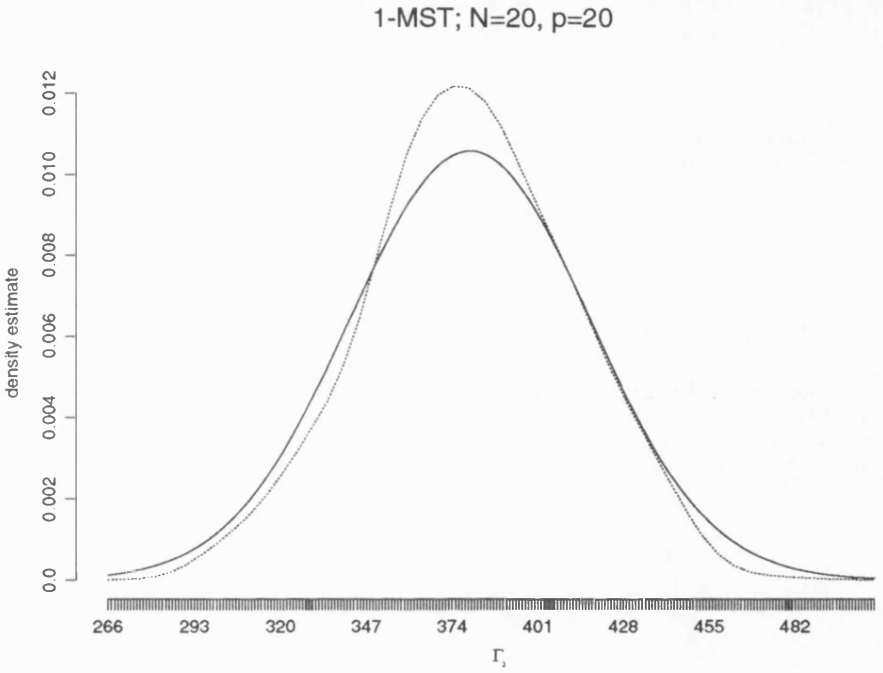


Figure 6-3: Γ_2 Approximations

Chapter 7

Case Studies

7.1 Introduction

We now illustrate the usefulness of the procedures studied in the previous chapters; we do so by analyzing three data sets. The purpose of these exercises is twofold. First, we demonstrate the tests in action in situations where it may not be entirely appropriate to use a parametric procedure. We also show the use of some of the graphs considered as a tool for exploratory data analysis, specifically as an aid to generate hypotheses.

We tested the multivariate normality hypothesis using the procedures developed by Mardia (1970). Mardia proposed the multivariate measures of skewness and kurtosis presented in the previous chapter and exploited the insights of two aspects of possible deviations from the multivariate normal distribution that may be gained from their use. We followed the algorithm given by Mardia and Zemroch (1975).

We should also point out that the overall performance of the tests, for shifts alternatives, can be improved if we transform the data in order to remove differences between the samples which may be caused by anything but the alternative hypothesis.

The example shown in Figure 7-1 should help to elucidate this. First suppose we have three samples from univariate distributions which differ only in location, as shown in Figure 7-1. The ranks based on the distances to the median should correspond to a batch of points from the middle sample and then to an almost alternate sequence of elements from the samples located at the extreme of the pooled data. This may cause

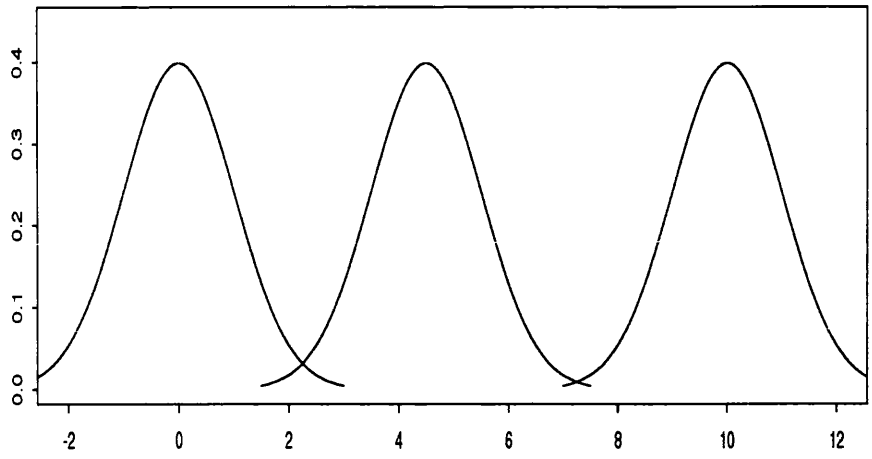


Figure 7-1: A situation in which rank tests may fail

the tests based on such ranks to reject, incorrectly, the hypothesis of equal variances. In order to correct the effect of different locations in tests based on radial ranks, we transform the data by substracting from each observation either the mean or the mediancentre of its sample.

In the next sections we analyze three data sets using the procedures studied in the previous chapters.

7.2 Reeve's Anteater Skulls Data

In this section we analyze the data originally published by Reeve (1941). The aim of his work was to assess the relevance of a classification of four subspecies of of the genus *Tamandua tetradactyla* (also known as anteaters bears) proposed by Allen in 1904. This taxonomy was based on differences in some lengths of parts of skulls. The data used with this purpose were measured on specimens procedent from samples collected all over America. Reeve considered the following three variables:

1. basal length, excluding the premaxilla

2. occipito-nasal length

3. greatest length of the nasals

All the measurements were made in milimeters. The skulls studied in Reeve's paper are from the subspecies *instabilis, chapadensis, chiriquensis,* and *mexicana.* Allen's hypothesis was that the skull measurements divided the four subspecies into two groups: *(instabilis, chapadensis)* and *(chiriquensis, mexicana).*

Parts of Reeve's data have been analyzed by Seal (1967), to illustrate the use of canonical correlation in multivariate analysis and by Mardia (1971) in the context of tests for multivariate normality and multivariate measures of skewness and kurtosis. Seber (1984) uses a subset of these data to illustrate the computations required by *MANOVA* tests. Gabriel (1968) also analyzed these data in an example of the use of his simultaneous tests procedure in multivariate analysis of variance. They are also briefly mentioned by Blackith and Reyment (1971) as an example of a data analysis in which two researchers (Seal and Reeve) did reach opposite conclusions analyzing the same data. We begin by presenting a summary of the data. Table 7.1 shows the descriptive statistics corresponding to these six samples. We worked with the logarithms to base 10 of these data in order to reduce them to a common order of magnitude. This transformation was performed by all the authors which have analyzed Reeve's data. An first inspection of Table 7.1 might suggest no differences for the mean vectors, and the opposite for the variance-covariance matrices.

A two dimensional representation of the complete data set appears in Figure 7-2. It was obtained using the first two principal components, which account for the (83% + 16%) of the total variation. It should be noted that the scales of the axes in the figure are different in order to show more clearly the differences among the samples.

Table 7.2 contains the values of those statistics and the significance levels associated with Mardia's approximations. The significance values obtained support the hypothesis of multivariate normality for all the samples. Of course, working with the logarithms of the data might be responsible for this agreement with the multivariate normal distribution. Nevertheless, for such small sample sizes the use of nonparametric procedures should be preferred whenever there is no additional information in favour

sample	n	mean	mediancentre	variance-covariance matrix		
instabilis	21	2.0539467	2.0553654	0.0002091	0.0001916	0.0003106
		2.0655858	2.0677169		0.0001902	0.0003106
		1.6208802	1.6221515			0.0008108
chapadensis 1	6	2.0967229	2.0899051	0.0007920	0.0008347	0.0010458
		2.0996730	2.0899051		0.0008961	0.0011404
		1.6252351	1.6127839			0.0014915
chapadensis 2	9	2.0905827	2.0935212	0.0004916	0.0004271	0.0003486
		2.0950380	2.0975828		0.0003973	0.0003239
		1.6244080	1.6333794			0.0005544
chapadensis 3	3	2.0991987	2.0995421			
		2.1015060	2.1003879		$n-1<p$	
		1.6432248	1.6501732			
chiriquensis	4	2.0924709	2.0969902	0.0000865	0.0000745	0.0001418
		2.1100612	2.1106829		0.0001249	0.0002962
		1.7025567	1.7025498			0.0008450
mexicana	5	2.0990497	2.1022776	0.0003103	0.0002738	0.0004547
		2.1070358	2.1113439		0.0002637	0.0004653
		1.6709639	1.6671054			0.0012275
All	48	2.0769000	2.0756833	0.0003245	0.0003040	0.0003914
		2.0856381	2.0838041		0.0003077	0.0004080
		1.6355063	1.6333162			0.0008608

Table 7.1: Reeve's data; descriptive statistics

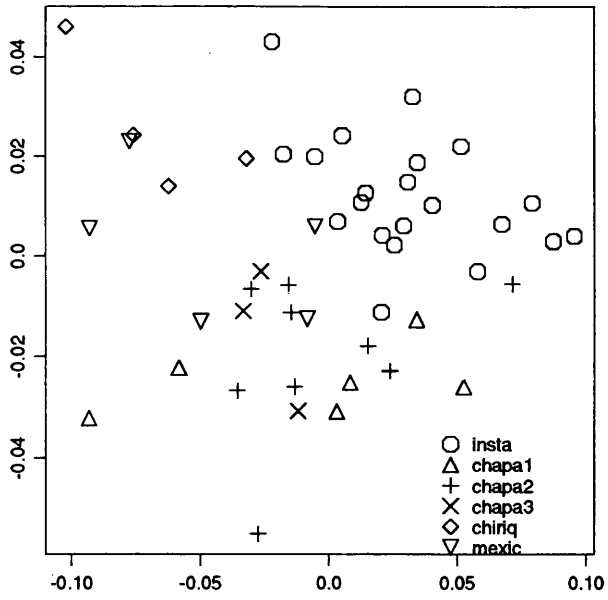


Figure 7-2: Reeve's data: two first principal components

of the claim of multivariate normality. Furthermore, Mardia's tests use approximations to their null distribution that may not be completely adequate for small sample sizes. Table 7.3 has the results obtained with the generalized runs tests. In order to obtain the corresponding significance value we fitted Pearson curves using the first four moments following the procedures described in a previous chapter. For all the graphs, the value of Γ_1 led to rejection of the hypothesis of homogeneity, with p-values smaller than 10^{-6}. As we mentioned, the runs tests are more sensitive to differences in location, so

sample	TV	$b(1,p)$	pv_1	$b(2,p)$	pv_2
instabilis	0.00115	3.09	0.37	14.43	0.40
chapadensis 1	0.00265	5.48	0.85	10.59	0.16
chapadensis 2	0.00128	4.37	0.77	11.93	0.20
chapadensis 3		$n-1<p$			
chiriquensis	0.00079	6.00	0.95	9.00	0.14
mexicana	0.00144	4.12	0.97	9.38	0.13
ALL	0.00269	1.85	0.14	15.59	0.65

Table 7.2: Reeve's data; generalized variance and multivariate skewness and kurtosis

graph	Γ_1	#(links)	μ	σ	β_1	β_2
1-*MST*	23	47	11.6667	2.6715	0.1168	2.9736
2-*MST*	49	94	23.3333	3.6711	0.2566	3.0679
3-*MST*	74	141	35.0000	4.5309	0.3412	3.1319
4-*MST*	95	188	46.6667	5.2035	0.3847	3.1234
5-*MST*	112	235	58.3333	5.8638	0.4200	3.0833
6-*MST*	130	282	70.0000	6.4101	0.4576	3.0068
1-*NNG*	18	36	8.9362	2.3383	0.1343	2.9810
2-*NNG*	34	65	16.1348	3.0275	0.2380	3.0607
3-*NNG*	45	93	23.0851	3.5526	0.2740	3.0884
4-*NNG*	60	120	29.7872	4.0122	0.3377	3.1177
5-*NNG*	74	148	36.7376	4.3987	0.3626	3.1144
6-*NNG*	90	182	45.1773	4.8499	0.3802	3.0840
1-*RNG*	27	56	13.9007	2.8557	0.1132	2.9774
2-*RNG*	62	145	35.9929	4.6104	0.3160	3.1242
1-*GG*	44	90	22.3404	3.6513	0.2062	3.0294
2-*GG*	108	249	61.8085	6.0623	0.5012	3.0072

Table 7.3: Reeve's data: multivariate runs tests results

these results should be interpreted accordingly to this fact. More information about the samples' distributions is obtained using multivariate ranks tests. Table 7.4 shows the significance values obtained with several parametric and multivariate ranks tests. The results have been arranged showing the location and scale alternatives versions of the tests. The first row corresponds to the *LRT* (Anderson (1984), Mardia et al. (1977)). As they are based on the assumptions of normality, we may feel confident in using them, given the results of Table 7.2. The numbers in parenthesis for the degree-1 tests are the number of leaves observed in each graph. All the tests lead to reject the equality of locations and to accept the hypothesis of homogeneity for the variance-covariance matrices.

So far we have rejected the hypothesis of homogeneous location for the six samples; however, it may be the case that all differ in location from one another, or there may exist homogeneous groupings of subsets of samples.

Gabriel (1968) devised a parametric procedure to construct simultaneous tests for the multivariate analysis of variance. As he points out, for this data set there are $2^6 - 6 - 1 = 57$ groups of samples and $2^3 - 1 = 7$ subsets of variables, which leads

location tests		scale tests	
	p-value		p-value
LRT	0.000000	*LRT*	0.544241
MMRST	0.000051		
MMMT	0.000001		
MMNST	0.000003	$(MMNST)^2$	0.135004
		deg1 *MST* (16)	0.307263
		deg1 *RNG* (7)	0.232719
K-W (diam)	0.000026	K-W (radial)	0.282186
NS (diam)	0.000047	NS (radial)	0.518769
Kiefer (diam)	0.000000	Kiefer (radial)	0.289722
S-S (diam)	0.003229	S-S (radial)	0.208869

Table 7.4: Reeve's data: parametric and multivariate ranks tests

to 399 null hypothesis that may be considered, plus hypotheses on linear combinations of subsets of variables and contrasts in 3 or more samples. Instead of following this rather cumbersome approach, which is based in asymptotic results and multivariate normality, we use the methods to examine relations between sample pairs outlined in Section 4.3. This is done in order to suggest further hypotheses to be tested. One possibility consists in comparing the observed and expected numbers of links between different samples. For example, Table 7.5 shows those numbers for the 1-*MST* and for the 1-*GG*.

The values of the χ^2 statistic were 89.14 and 158.73, which are higly significant for 14 degrees of freedom. It is apparent from Table 7.5 that *instabilis* is the most isolated sample, while *mexicana* might be suspected of being too near to *chapadensis* and *chiriquensis*. This table also suggests that *chiriquensis* does differ from all the other samples, except possibly from *mexicana*. At this stage, some questions may arise.

- Are the three samples of *chapadensis* homogeneous?

  Considering only these samples, we used all the tests, except the *LRT* ones. With the runs tests, we overwhelmingly accepted the null hypothesis; all the significance values were larger than 0.25. The same happened for the rank tests, with two exceptions: the Kruskal-Wallis location test and the Normal Scores

samples			1-*MST* OBS	1-*MST* EXP	1-*GG* OBS	1-*GG* EXP
insta	-	*chap1*	1	4.80	4	10.16
insta	-	*chap2*	4	7.20	6	15.24
insta	-	*chap3*	0	2.40	1	5.08
insta	-	*chiri*	2	3.20	2	6.77
insta	-	*mexic*	1	4.00	2	8.47
chap1	-	*chap2*	3	2.05	4	4.35
chap1	-	*chap3*	1	0.68	1	1.45
chap1	-	*chiri*	0	0.91	0	1.93
chap1	-	*mexic*	1	1.14	5	2.42
chap2	-	*chap3*	1	1.02	9	2.17
chap2	-	*chiri*	0	1.37	0	2.90
chap2	-	*mexic*	4	1.71	6	3.63
chap3	-	*chiri*	1	0.45	1	0.96
chap3	-	*mexic*	1	0.57	1	1.21
chiri	-	*mexic*	2	0.76	4	1.61

Table 7.5: Reeve's data: number of links from different samples

location test yielded the significance values 0.018 and 0.021, respectively. Both tests also produced the smallest p-values for the scale alternative (0.177 and 0.181, respectively). On the other hand, all the test statistics whose distributions were approximated by matching the first four moments or with a large number of sample identities permutations (Scholz-Stephens test and runs tests) consistently accepted the null hypotheses of equal locations and scales. The Kruskal-Wallis and the Normal Scores tests p-values are based on an asymptotic approximation, and we may think that the relatively small total sample size ($N = 18$) was not enough to produce good results for those two statistics or that averaging the within-sample ranks is not the most efficient procedure for analyzing very small sample sizes. A similar example can be seen in Scholz and Stephens (1987). We also tested the homogeneity of the four subspecies, joining the three samples of *chapadensis*. All the nonparametric tests led to accept the equality of variance-covariance matrices and to reject the homogeneity of locations.

- Does *instabilis* differ from the other three samples?

  The answer is *yes*. We tested *instabilis* pairwise against the other three subspecies always resulting in rejection of the homogeneity of locations hypothesis and accepting the equality of variance-covariance matrices.

- Do *chapadensis* and *chiriquensis* have different locations?

  Yes. All the tests categorically rejected the hypothesis of homogeneity of locations, while accepting the equality of variance-covariance matrices. This result was advanced by the comparisons observed in Table 7.2.

- Does *mexicana* significantly differ from *chapadensis* and *chiriquensis*?

  The answer is *no* and *no*. It is worth noting that when comparing *mexicana* and *chapadensis* (with n=18), the parametric tests for location produced p-values of about 0.02, while the significance levels corresponding to most of the nonparametric tests were around 0.20. An exception was the runs test based on the 1-*NNG* which yielded a significance value of 0.046 calculated using all the possible permutations (126) of sample labels over the pooled data. All the corresponding tests accepted unequivocally the homogeneity of the variance-covariance matrices. For the comparison between *mexicana* and *chiriquensis*, all the tests indicated that the data from both subspecies were homogeneous in an even clearer way than the one observed when comparing *mexicana* and *chapadensis*.

Seal (1967) concluded that the six subspecies were different from one another. She based this claim on a a graph of the projection of the sample means and a rough estimate of their confidence intervals on the first two canonical variates. However, as Gabriel (1968) noted, this conclusion is not based on any statistical significance level and should not be taken as a confirmatory result. Reeve (1941) concluded that Allen's classification of the four subspecies into two groups was erroneous. He stated that "only *chapadensis* and *mexicana* have any claim to be considered distinct subspecies on the basis of skull proportions". We have seen that this is not correct, as *chiriquensis* and *instabilis* are different from each other and also from the first two subspecies.

As we noted, *mexicana* could be joined with either *chiriquensis* or *chapadensis*. Only an analysis with more data would clear up this point.

7.3 Lubischew's Beetle Data

Lubischew (1962) applied discriminant analysis to a total of 74 specimens from three different species of male flea-beetles of the genus *Chaetocnema*; the species were *concinna* ($n_1 = 21$), *heikertingeri* ($n_2 = 31$) and *heptapotamica* ($n_3 = 22$). The variables considered by Lubischew were

1. width of the first joint of the first tarsus, in microns (sum for both tarsi)
2. width of the second joint of the first tarsus, in microns (sum for both tarsi)
3. the maximal width of the head between the external edges of the eyes, in units of 0.01mm
4. the maximal width of the aedeagus in the fore part, in microns
5. the from angle of the aedeagus, in units of 7.5^0
6. the aedeagus width from side in microns

This data set has also been analyzed by Jones and Sibson (1987) and by Taylor (1987). These authors have noted that the samples are well differentiated and proposed several classification rules in order to provide easier ways of distinguishing the procedence of new data from the three species.

We begin by showing some descriptive statistics of the data. From Table 7.6, one might think that the samples differ in location and in variance-covariance matrix.

The data exhibit some evidence of departures from a multivariate normal distribution, as the values for $b(2, p)$ seem to be significant for the first and the pooled samples, and so, it may be appropriate to use nonparametric procedures. It is worth noting that Table 7.7 shows that the two observed significant values of $b(2, p)$ were due to an excess of "flatness" in the distribution. An asterisk indicates that the value is significant in that direction.

sample	*n*	mean	medcen		variance-covariance matrix				
conc.	21	183.09	183.31	147.49	66.63	18.52	15.08	-5.21	14.21
		129.61	130.36		51.24	11.54	2.47	-1.81	3.09
		51.23	51.299			4.99	5.85	-0.52	5.48
		146.19	146.69				31.66	-0.97	15.62
		14.09	14.029					0.79	-1.98
		104.85	105.29						38.22
heike.	31	201.00	199.92	222.13	63.40	22.60	30.36	4.36	29.46
		119.32	119.41		44.15	7.91	11.81	0.33	11.46
		48.87	48.67			5.51	5.68	0.01	4.23
		124.64	124.59				21.36	-0.32	11.70
		14.29	14.18					1.21	1.26
		81.00	81.10						79.73
heptap.	22	138.22	137.55	87.32	44.55	20.52	19.17	-0.73	15.28
		125.09	124.24		73.03	15.70	14.02	-0.38	21.22
		51.59	51.61			8.06	8.21	-0.29	4.96
		138.27	138.14				17.16	-0.50	7.92
		10.09	10.21					0.94	0.27
		106.59	105.96						34.25
ALL	74	177.25	180.34	865.09	6.57	-7.74	-101.97	49.22	-240.53
		123.95	124.46		71.92	15.71	49.29	-2.21	59.23
		50.35	50.20			7.57	16.88	-1.84	20.31
		134.81	135.61				107.14	-5.56	116.16
		12.98	13.45					4.58	-14.66
		95.37	95.64						204.62

Table 7.6: Lubischew's data; descriptive statistics

We also transformed this data using their natural logarithms, in an attempt of having measurements of the same order of magnitude for each variable. This transformation did not correct the lack of normality; moreover, it had the opposite effect, producing a significantly skewed distribution for the pooled data and distributions differing from the multivariate normal by both skewness and kurtosis for samples 1 and 2. A two-

sample	TV	$b(1,p)$	pv_1	$b(2,p)$	pv_2	
concinna	261.344	11.1546	0.9587	40.8396	0.9638	*
heikertingeri	362.052	9.6985	0.5963	43.4283	0.9141	
heptapotamica	210.750	13.5231	0.7148	43.3478	0.8783	
ALL	1243.914	4.7106	0.3979	43.6899	0.9824	*

Table 7.7: Lubischew's data; generalized variance and multivariate skewness and kurtosis

dimensional representation of the standarized data obtained using the *planing* method of Friedman and Rafksy (1981) appears in Figure 7-3. We obtained the corresponding 1-*MST* with the Euclidean distance matrix calculated over the standarized data.

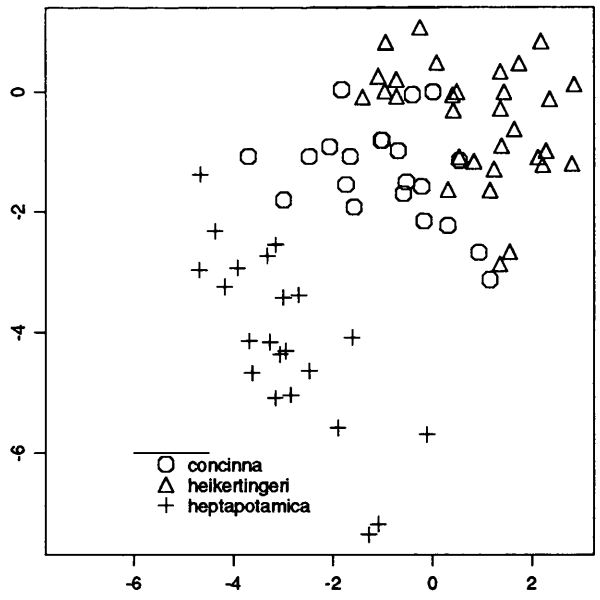


Figure 7-3: Lubischew's data: *planing* representation

Table 7.8 has the p-values for these data. All the tests coincide in rejecting the equality

location tests		scale tests	
	p-value		p-value
LRT	0.000000	*LRT*	0.103509
MMRST	0.000000		
MMMT	0.000000		
MMNST	0.000000	$(MMNST)^2$	0.002013
		deg1 *MST* (29)	0.663810
		deg1 *RNG* (16)	0.204100
K-W (diam)	0.000000	K-W (radial)	0.000005
NS (diam)	0.000000	NS (radial)	0.000005
Kiefer (diam)	0.001130	Kiefer (radial)	0.001418
S-S (diam)	0.010000	S-S (radial)	0.001000

Table 7.8: Lubischew's data: parametric and multivariate ranks tests

samples	1-*MST*		1-*GG*	
	OBS	EXP	OBS	EXP
concinna - *heikertingeri*	1	17.59	19	81.96
concinna - *heptapotamica*	1	12.48	29	57.81
heikertingeri - *heptapotamica*	0	18.42	21	85.34

Table 7.9: Lubischew's data: number of links from different samples

of locations. For the homogeneity of variance-covariance matrices, the parametric and the degree-1 tests accept the null hypothesis, while the rest of the nonparametric tests reject it. All the multivariate runs based tests 1 rejected the hypothesis of homogeneity with significant levels smaller than 10^{-6}. All the graphs were obtained using Euclidean distances calculated over the standarized . Table 7.9 gives the observed and expected numbers of links from different samples; it shows the pairwise gaps existing between the three species.

As we saw, there is some evidence of non normality indicated by a significant value of the multivariate kurtosis coefficient $b(2,p)$. Mardia (1970) has shown that the *LRT* for equality of variance-covariance matrices are sensitive to non normality and that such sensitivity is indicated by $b(2,p)$, while the *LRT* for location alternatives are sensitive to significant values of the multivariate skewness coefficient. This means that we cannot rely strongly on the acceptance of the equality of variance covariance

	location tests			scale tests	
	mean	medcentre		mean	medcentre
LRT	0.995470	0.992678	*LRT*	0.103509	0.103509
MMRST	1.000000	1.000000			
MMMT	0.998719	0.999059			
MMNST	0.999988	0.999987	$(MMNST)^2$	0.012551	0.023505
			deg1 *MST*	0.095022	0.291839
			deg1 *RNG*	0.250073	0.411946
K-W (diam)	0.090070	0.735333	K-W (radial)	0.986385	0.796501
NS (diam)	0.102158	0.825375	NS (radial)	0.985027	0.796501
Kiefer (diam)	0.498543	0.605994	Kiefer (radial)	0.105374	0.194204
S-S (diam)	0.250000	0.250000	S-S (radial)	0.215323	0.250000

Table 7.10: Lubischew's data: tests with centered samples

matrices pointed out by the *LRT*. On the other hand, it appears that the degree-1 tests have a better performance for distributions with values of the kurtosis coefficient which are significantly above the mean of the null distribution of $b(2,p)$ (e.g. lognormal or Cauchy).

As these data might be an instance of the situation exemplified in Figure 7-1 a, we calculated the statistics based on the radial rankings for the 1-*MST* and the 1-*RNG* obtained by substracting the corresponding sample mean or mediancentre from the data points. The results are presented in Table 7.10. We now have that, except for $(MMNST)^2$, all the tests support the hypothesis of equality of variance covariance matrix. The numbers of leaves for the 1-*MST*s for the substracted mean and the mediancentre were 27 and 28; the corresponding numbers for the 1-*RNG*s were 11 and 9. It should also be noted that without any exception, the tests based on runs accepted by far the null hypothesis for the data centered by the means and by the mediancentres. As these tests are conceived as tests for complete homogeneity and the variation due to different locations have been removed, we can conclude that these results also point towards accepting the null hypothesis against the alternative of different scales.

NORTHERN UTOAZTECAN		SOUTHERN UTOAZTECAN	
NUMIC	Mono Comanche Southern Paiute	TEPIMAN	Papago Northern Tepehuan Southern Tepehuan
TAKIC	Luiseño Cahuilla Serrano	TARACAHITAN	Tarahumara Guarijio Tubar Yaqui Mayo
Hopi		CORACHOL	Cora Huichol
Tübatulabal		AZTECAN	Nahuatl of Zacapoaxtla

Table 7.11: Utoaztecan languages

7.4 Utoaztecan Languages

In this section we give an example of the use of the correlation coefficients studied in Chapter 6.

The Utoaztecan (UA) languages are still spoken from Utah to El Salvador and form a major family of the Amerindian languages of North and Central America. UA languages have been an important research area in American linguistics since the arrival of the Spaniards. However, not much is known about the internal relationships within the family.

The data analyzed in this section were initially collected from a great variety of sources by Professor Leopoldo Valiñas Coalla, of the Instituto de Investigaciones Antropológicas in Mexico City. They consist of lexical and phonological evidence for 19 UA languages, whose names, groups, and subgroups appear in Table 7.11 The data appear in the papers by Cortina and Valiñas (1989, 1990). Hopi and Tübatulabal are thought to be isolated languages.

The lexical data consisted of Swadesh's list of 100 words for each language. These words represent basic vocabulary which is regarded as being resistant to changes induced by time or cultural contacts. Some examples are *I, you, woman, man, one, two, blood, nose,* and *eye*. A density table, considering the cognated and non-cognated forms for pairs of languages was constructed from these words. The lexical dissimilarity between any two languages was defined as one minus their cognate density. This insures that the more similar in lexicon are two languages, the smaller their lexical dissimilarity is.

A measure of phonological dissimilarity was constructed from two tables which contained the reflexes of the actual phonological systems for each language into some of the protophonems reconstructed for the UA languages. The methodology for constructing such protophonems is thoroughly discussed by Arlotto (1972).

Those tables registered the changes in the observed phonems in the beginning of words (for 17 protophonems) and in between vowels (for 13 protophonems) found for each language. Our interest in this section is to examine how close the relations of similarity found for the phonological data resemble those observed within the lexical data for the same languages.

We can use cluster analysis, *K*-means, multidimensional scaling and other methods which are adequate tools for analyzing distance matrices. However, they do not directly provide us with results which are susceptible of statistical inference. Also, a better understanding of the data might be achieved by working at a more basic level. This means carefully examining the pairs of languages that lie nearby each other without the structural interference inherent to clustering methods and also without having the more global vision imposed by multidimensional scaling plots.

In Figures 7-4 and 7-5 we present the configurations obtained with the *planing* algorithm for the 1-*MST* of Friedman and Rafsky (1981), together with the corresponding 1-*RNG*s. These graphs reveal some features present in the dissimilarity matrices which may be of interest for the linguist.

Following the arguments of Lefkovitch (1984), we worked with the 1-*RNG*, as it is the graph which would be more robust to possible deviations from non-Euclidean

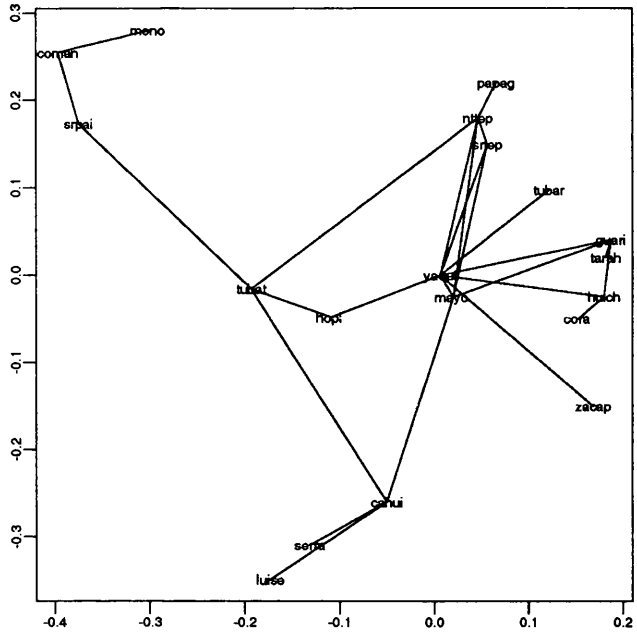


Figure 7-4: UA data: RNG for lexical data

distances. Another point in favour of this graph is that, as Toussaint (1980) remarked, the 1-*RNG* shows a more "rounded" vision of the important distances than those obtained with the 1-*NNG* or the 1-*MST*.

We found 11 common edges in the *RNG*s for the phonological (25 edges) and the lexical (21 edges) data. Thus, $\Gamma_1 = 11$. Using expressions (6.2) and (6.3), we found that the mean and standard deviation of the null distribution of Γ_1 conditioned on the observed *RNG*s were 3.07 and 1.49, respectively. So the normal approximation overwhelmingly signaled towards rejecting the hypothesis of no correlation.

The significance of this results lies in the fact that the phonological data were gathered from many sources, and thus, if there were wide divergences between them and the lexical evidence, we would treat them suspiciously.

We calculated Γ_1 for 10000 permutations over the edges of the *RNG* corresponding to the phonological data. The result was identical to that obtained with the normal approximation. A density estimate of the permutational distribution and the normal approximation appear in Figure 7-6. We used a Gaussian kernel and chose the smoothing parameter using the optimum value proposed by Silverman (1986; §3.4).

The prediction coefficient Γ_2 also lead unequivocally to reject the null hypothesis.

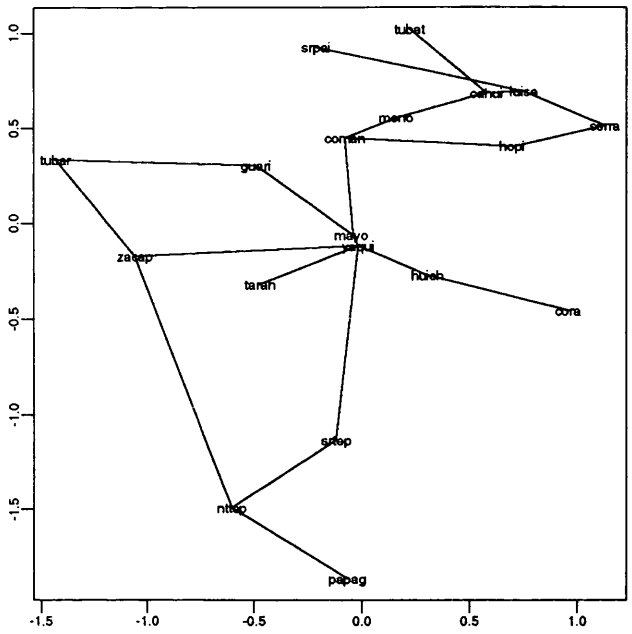


Figure 7-5: UA data: RNG for phonological data

We had $\Gamma_2 = 223$. The mean and the standard deviation, calculated according to expressions (6.7) and (6.8) are 399 and 37.82, respectively, thus producing a highly significant result. This result was the same when comparing the observed value of Γ_2 with its permutational null distribution, approximated by 10000 values. A density estimate of this distribution and the normal approximation are shown in Figure 7-7. Table 7.12 presents the descriptive statistics for the permutational approximations of Γ_1 and $\Gamma 2$ based on 10000 simulations. The permutational distribution of Γ_2 seems to be very close to a normal distribution, while the one of Γ_1 could be represented better by a Pearson Type VI distribution.

	$\bar{x}$	s	b_1	b_2	κ
Γ_1	3.0701	1.4908	0.0545	3.0946	1.5997
Γ_2	399.59	32.0243	0.0009	3.0191	-0.0166

Table 7.12: UA data: Skewness and Kurtosis measures for Γ_1 and Γ_2

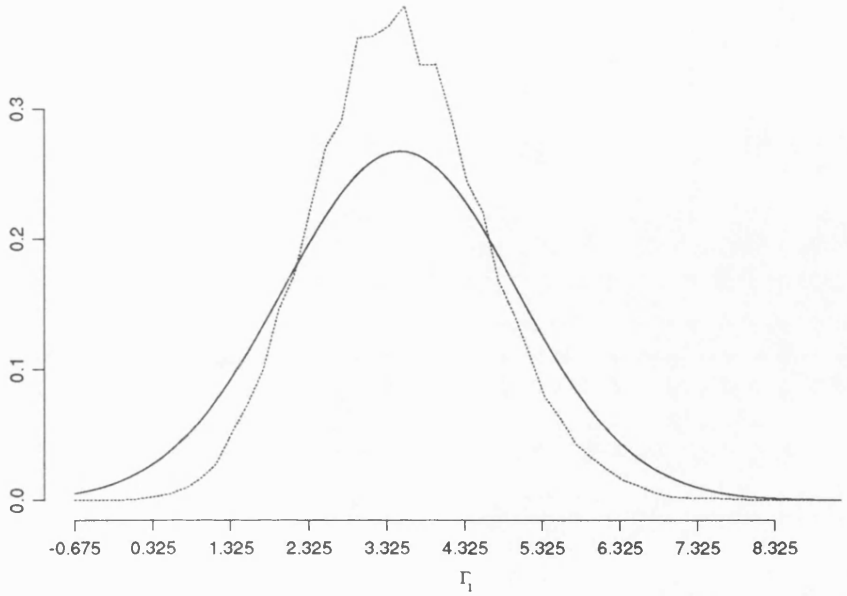


Figure 7-6: UA data: Permutational and Normal approximations for Γ_1

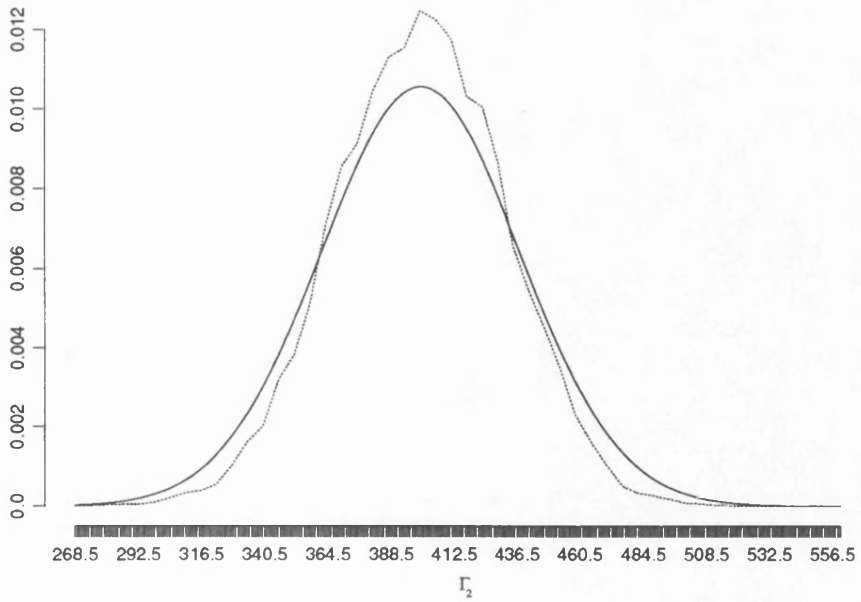


Figure 7-7: UA data: Permutational and Normal approximations for Γ_2

Chapter 8

Final Remarks

8.1 Conclusion

The original aim of this thesis was to provide a battery of procedures to analyze multivariate observations. Our motivation was based on practical situations in which the sample sizes are too small for the researcher to decide if the assumptions required by parametric tests of hypothesis hold. Therefore, the investigation started off focusing on the areas of nonparametric tests and their small samples approximations. We noticed that the mathematical machinery employed to construct test statistics has a use of its own in initial analysis of multivariate observations. The examples presented in the previous chapter illustrate this point.

Making inferences in multivariate analysis is not an easy task. Even when we could establish in a direct manner the validity of the procedures used, we usually would have to choose amongst several statistics for a particular situation. Guidelines are, when available, rather vague. We fully agree with Krzanowski (1988, §8.4) when he affirms that hypothesis testing in the multivariate case should be used in the "informal inference" sense also advocated by Chatfield (1985). By this we understand that the significance levels obtained for any practical application involving multivariate observations should be considered as markers of the presence of particular features of the data, rather than in a rigid decision-making framework.

As Sibson (1984) pointed out, multivariate procedures which are strongly based on

normal theory and which mainly take into account the analysis of mean vectors and variance-covariance matrices are not always successful for describing the complex relations within the observations. We have surveyed a variety of strategies which attempt to extract information from the data in a more flexible way. We hope that this work will draw attention towards procedures which might be a useful alternative to those of classical multivariate analysis and which are a substantive addition to the expanding field of exploratory multivariate data analysis techniques.

In the next sections we briefly address some research lines which are related to the main topics covered in the previous chapters. We have already done some work in several of these areas, which should be continued in due course.

8.2 Planing

As we mentioned in the first chapter, *planing* is a mapping technique based on preserving a few key distances from the original configuration. It is based on a triangulation method due to Lee et al. (1977) and attempts to preserve the distances corresponding to the edges of the 1-*MST*. Friedman and Rafsky (1981) introduced this technique and gave some examples of its performance on several multivariate data sets. It is easy to construct versions of the original *planing* procedure in order to obtain configurations in 1 and 3 dimensions. For the latter case, there are many arbitrary decisions to be made, and, as a result, the number of possible configurations increases. More work is necessary with these ideas in order to provide flexible algorithms.

Siedlecki et al. (1988a) reviewed a great number of mapping techniques. The same authors (1988b) discussed several examples of these techniques applied to problems in pattern recognition. It would be interesting to compare some versions of *planing* with the mapping techniques presented by those authors.

Planing is a very fast method indeed. Friedman and Rafsky (1981) reported experiments which successfully dealt with thousands of multivariate points. It can be very useful as a tool geared towards providing a reasonable first approximation for projection methods which require an initial configuration.

Another interesting possibility is to use the *ET* instead of the 1-*MST* in order to construct low dimensional representations of multivariate data. This enables us to construct configurations which would emphasize the ordered list of distances from the particular individual chosen as the root of the *ET*.

8.3 Nonparametric Tests

While searching for nonparametric tests which could be used jointly with the multivariate ranking procedures we noticed that although a great variety of tests are available from the literature, very little is known about the small sample properties of many of them. For instance, for Kiefer (1959) or Scholz-Stephens (1987) *K*-sample tests some elegant asymptotic approximations are available. However, virtually nothing is known about their power and efficiency properties or the way they would react with different ranks. Whaley and Quade (1987) proposed to work with links defined by a threshold distance and found that the power of the runs test was improved by that procedure. It would be interesting to proceed in a similar way using the relative neighbourhood graphs presented by Urquhart (1980, 1982) in order to construct generalizations of the univariate runs test.

We also noted that the multivariate rank procedures studied by Puri and Sen (1971) seem to have good power properties. For location alternatives, we observed that *MMNST* is apparently better than the *MMRST* and *MMMT*.

For scale alternatives, we only considered a test based on the squares of the normal scores. Of course, there are many other possibilities for constructing multivariate rank tests which may be more suitable in different situations. It would be useful to obtain more detailed guidelines in order to select amongst these tests and other possibilities based on Puri and Sen's approach.

8.4 Multivariate Ranks

In Chapter 3 we performed a very limited study comparing one of the Friedman-Rafsky multivariate ranking procedures with the ordering obtained by mapping the multivariate observations along their first principal component. We could observe that, although both procedures' performance deteriorated as the dimensionality of the data increased, the diameter ranking based on the 1-*MST* did a better job in terms of ranking together points which were near to each other in the multivariate space. It would be interesting to compare in depth the *MST*-based ranking procedures with other methods. Barnett (1976) presented a review of procedures for ordering multivariate data which could be used as a starting point for such a task.

8.5 Computational Geometry

Another line of work suggested by this research is in computational geometry. Three questions which one might think of are:

1. How well does a sequence of orthogonal minimum spanning trees approximate a Dirichlet Tessellation in the plane? For instance, to what extent would such an approximation satisfy some properties enjoyed by the *DT*, e.g. equiangularity as studied by Sibson (1978)?

2. Is it possible to approximate a p-dimensional *DT* using sequences of *MST*s, *RNG*s or *GG*s?

3. How can we extend Supowit's (1983) fast algorithms for planar *RNG*s to higher dimensions?

8.6 Geometrical Probability

A very interesting field which has not received much attention is the characterization of some statistical properties of graphs for random patterns.

Some questions of interest would concern the distributions of the number of circuits (or any other pattern that might be formed with a few edges), of the node degrees, and of the total edge length. Several papers (e.g. Bearwood et al. (1959), Roberts (1968, 1969), Steele (1980, 1988), Bertsimas and van Ryzin (1990)) have been devoted to establishing estimates and bounds for the expected length of the 1-*MST* of a set of N points in p dimensions and for the asymptotic distribution of the number of leaves in the same graph. However, there is still much to be done about these problems. For instance, it would be useful to conduct an investigation about the values of the asymptotic constants depending only on p needed to calculate the expected length of the graphs. Such values could be used later as a basis for constructing uniformity tests for point patterns.

It is easy to obtain the moments of the distribution of the total lengths for the *ET*s, *NNG*s, *RNG*s and *GG*s constructed for randomly distributed points. As the formulae necessary to do so are based on the corresponding region of influence that defines each graph, it seems that generalizing these results for graphs in Urquhart's families of relative neighbourhood graphs is a straightforward problem.

8.7 Spatial Statistics

Another possibility of research arising from the thesis is in using orthogonal *MST*s or generalized relative neighbourhood graphs in order to construct tests for spatial randomness as those studied by Ripley (1979, 1981).

Mead (1966), Sibson (1981) and Upton and Fingleton (1985, §5.4), amongst others, have developed some statistics for testing spatial randomness using the *DT* for points in the plane. The idea is to calculate the values of several characteristics of the *DT* for a spatial pattern and to compare them with the distributions presented by Hinde and Miles (1980) and Quine and Watson (1984) for planar Poisson processes.

It would be very interesting to find out more about the distributions of similar characteristics for other graphs and to compare their performances with tests based on the *DT*.

8.8 Multivariate Outliers

Rohlf (1975) introduced a test for identifying multivariate outliers based on the longest edge of the 1-*MST* constructed on the Euclidean distance of the data point. This test generalizes the one proposed by Dixon (1950) for the univariate case. Rohlf (1975) discussed an approximation to the distribution of the longest edge of the 1-*MST* for multivariate Normal data. It would be interesting to enhance his results for other multivariate distributions, perhaps using different graphs to do so.

References

ANDERSON, TW (1984) *Introduction to Multivariate Analysis.* (2nd edition). John Wiley and sons: New York.

ANDERSON, TW AND DARLING, DA (1954) A test of goodness of fit. *J. Amer. Statist. Ass.* **49** 765-769.

ARLOTTO, A (1972) *Introduction to Historical Linguistics.* Houghton Mifflin Company: Boston.

BARNETT, V. (1976) The ordering of multivariate data. *J. Roy. Statist. Soc. A* **139**, 318-354.

BARTON, FN AND DAVID, FN (1966) The random intersection of two graphs. In *Research Papers in Statistics.* (David, FN and J Neyman, eds.), pp. 445-459. John Wiley and sons: New York.

BEARWOOD, J HALTON, JH AND HAMMERSLEY JM (1959) The shortest path through many points. *Proc. Camb. Phil. Soc.* **55**, 299-327.

BEDALL, FK AND ZIMMERMANN, H (1978) Algorithm AS 143. The Mediancentre. *Appl. Stat.* **27**, 325-328.

BENTLEY, JL (1978) Fast algorithm for constructing minimal spanning trees in coordinate spaces. *IEEE Trans. Comp.* **C27**, 97-105.

BERRY, KJ (1982) Algorithm AS 179: Enumeration of all permutations of multi-sets with fixed repetition numbers. *Appl. Stat.* **31**, 169-173.

BERRY, KJ AND MIELKE, PW (1983) Computation of finite population parameters and approximate probability values for multi-response permutation procedures (MRPP). *Comm. Stat. Theory and Methods* , **12**, 83-107.

BERRY, KJ AND MIELKE, PW (1984) Computation of exact probability values for multi-response permutation procedures (MRPP). *Comm. Stat. Theory and Methods* , **13**, 417-432.

BERRY, KJ, MIELKE, PW AND WONG, RKW (1986) Approximate MRPP p-values obtained from four exact moments. *Comm. Stat. Theory and Methods* , **15**, 581-589.

BERTSIMAS, DJ AND VAN RYZIN, G (1990) An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability. *Oper. Res. Lett.* **9**, 223-231.

BIRNBAUM, ZW AND HALL, RA (1960) Small sample distributions for multisample statistics of the Smirnov type. *Ann. Math. Stat.* **31** 710-720.

BLACKITH, RE AND REYMENT, RA (1971) *Multivariate Morphometrics* Academic Press: London.

BLUMENTHAL, S (1963) The asymptotic normality of two test statistics associated with the two sample problem. *Ann. Math. Stat.* **34** 1513-1523.

BOOTS, BN AND MURDOCH, DJ (1983) The spatial arrangement of random Voronoi polygons. *Comput. Geosc.* **9**, 351-365.

BOWYER, A (1981) Computing Dirichlet tessellations. *Comput. J.* **24**, 162-167.

BROHET, CR (1984) Computer interpretation of pediatric orthogonal electrocardiograms: statistical and deterministic classification methods. *Circulation* **70**, 255-263.

CAPON, J (1965) On the asymptotic efficiency of the Kolmogorov-Smirnov test. *J. Amer. Statist. Ass.* **60**, 843-853.

CHASE, GR AND BULGREN, WG (1971) A Monte Carlo investigation of the robustness of T^2. *J. Amer. Statist. Ass.* **66**, 499-502.

CHATFIELD, C (1985) The initial analysis of data (with discussion). *J. Roy. Statist. Soc. A* **148**, 214-253.

CHATFIELD, C AND COLLINS, AJ (1980) *Introduction to Multivariate Analysis.* Chapman and Hall, London.

CHUNG, JH AND FRASER, DAS (1958) Randomization tests for a multivariate two-sample problem. *J. Amer. Statist. Ass.* **53**, 729-735.

CLEVELAND, WS (1979) Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Ass.*, **74**, 829-836.

CLIFF, AD AND ORD, JK (1980) *Spatial Processes: Models and Applications.* Pion: London.

CONSTANZO, CM, HUBERT, LJ AND GOLLEDGE, RG (1983) A higher moment for spatial statistics. *Geogr. Anal.*, **15** 347-351.

CONOVER, WJ (1965) Several *k*-sample Kolmogorov-Smirnov tests. *Ann. Math. Stat.* **36**, 1019-1026.

CONOVER, WJ (1971) *Practical Nonparametric Statistics.* John Wiley and sons: New York.

CORTINA BORJA, MJF AND VALIÑAS, L (1989) Some remarks on Utoaztecan classification. *Int. J. Amer. Ling.* **55**, 214-240.

CORTINA BORJA, MJF AND VALIÑAS, L (1990) Análisis léxico y fonológico de la familia yutoazteca. *Memorias del I Coloquio Swadesh*, Instituto de Investigaciones Antropológicas: México.

COX, DR AND HINKLEY, DV (1974) *Theoretical Statistics.* Chapman and Hall: London.

COX, TF (1981) Reflexive nearest neighbours. *Biometrics* **37** 367-369.

D'AGOSTINO, RB, BELANGER, A AND D'AGOSTINO, RB (1990) A suggestion for using powerful and informative tests of normality. *Amer. Statistician* **44**, 316-321.

DANIELS, HE (1944) The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33**, 120-135.

DAVID FN AND BARTON, DE (1960) *Combinatorial Chance.* Griffin: London.

DAVIS, AW (1980) On the effects of moderate multivariate nonnormality on Wilk's likelihood ratio criterion. *Biometrika* **67**, 419-427.

DAVIS, CS AND STEPHENS, MA (1983) Approximate percentage points using Pearson curves. *Appl. Stat.* **32**, 322-324.

DIGGLE, PJ (1983) *Statistical Analysis of Spatial Point Patterns* Academic Press, London.

DIXON, WJ (1950) Analysis of extreme values. *Ann. Math. Stat.* **21**, 488-506.

DU TOIT, SHC, STEYN, AGW AND STUMPF, RH (1986), *Graphical Exploratory Data Analysis.* Springer Verlag: New York.

ELDERTON, WP AND JOHNSON, NL (1969) *Systems of Frequency Curves.* Cambridge University Press: Cambridge.

EVERITT, BS (1979) A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample T^2 tests. *J. Amer. Statist. Soc.*, **74**, 48-51.

FRIEDMAN, JH, BASKETT, F AND SHUSTEK, LJ (1975) An algorithm for finding nearest neighbours. *IEEE Trans. Comput.* **C-24** 1000-1006.

FRIEDMAN, JH AND RAFSKY, LC (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Stat.* **7** 697-717.

FRIEDMAN, JH AND RAFSKY, LC (1981) Graphs for the multivariate two-sample problem. *J. Amer. Statist. Ass.* **76**, 277-287.

FRIEDMAN, JH AND RAFSKY, LC (1983) Graph-theoretic measures of multivariate association and prediction. *Ann. Stat.* **11** 377-391.

GABRIEL, KR (1968). Simultaenous test procedures in multivariate analysis of variance. *Biometrika* **55**, 289-504.

GABRIEL, KR AND SOKAL, RR (1969). A new statistical approach to geographic variation analysis. *Syst. Zool.*, **18**, 259-278.

GILBERT, EN (1965) Random minimal trees. *SIAM J. Appl. Math.* **13**, 376-387.

GREEN, PJ (1981) Peeling bivariate data. In *Interpreting Multivariate Data* (V. Barnett, ed.), pp. 3-20, John Wiley and sons: New York.

GREEN, PJ AND SILVERMAN, BW (1979) Constructing the convex hull of a set of points in the plane. Comput. J. **22**, 262-266.

HENZE, N (1988) A multivariate two-sample test based on the number of nearest neighbour type coincidences. *Ann. Stat.* **16** 772-783.

HINDE, AL AND MILES, RE (1980) Monte Carlo estimates of the distributions of the random polygons of the Voronoi Tessellation with respect to a Poisson Process. *J. Statist. Comp. Sim.* **10**, 205-223.

HOPE, ACA (1968) A simplified Monte Carlo significance test procedure. *J. Roy. Statist. Soc. B*,**30**, 582-598.

HUBERT, LJ (1974) Some applications of graph theory to clustering. *Psychometrika*, **39**, 283-309.

JANSON, S (1986) Random trees in a graph and trees in random graphs. *Math. Proc. Camb. Phil. Soc.*, **100**, 319-330.

JARDINE, N AND SIBSON, R (1971) *Mathematical Taxonomy.* John Wiley and sons: New York.

JONES, MC AND SIBSON, R (1987) What is projection pursuit? *J. Roy. Statist. Soc. A*, **150**, 1-37.

JÓŹIK, A (1983) A method for solving the n-dimensional convex hull problem. *Patt. Recogn. Lett.* **2**, 23-25.

KARLIN, S CAMERON, EC AND CAKRABORTY, R (1983) Misconceptions ins trials of structured exploratory data analysis. *Amer. J. Hum. Genet.* **35**, 695-673.

KENDALL, MG (1962) *Rank Correlation Methods.* Griffin: London.

KENDALL, MG, STUART, A AND ORD, JK (1987) *Kendall's Advanced Theory of Statistics.* Charles Griffin and co: London.

KIEFER, J (1959) K sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Ann. Math. Stat.* **30**, 420-447.

KLOTZ, J (1964) On the normal scores two sample rank test. *J. Amer. Statist. Ass.* **59**, 652-664.

KRUSKAL, JB (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.* **7** 48-50.

KRUSKAL, JB (1977) Multidimensional scaling and other methods for discovering structure. In *Statistical Methods for Digital Computers* (A. Ralston Einstein and H. Wilf, eds). John Wiley and sons: New York.

KRZANOWSKI, WJ (1988) *Principles of Multivariate Analysis. A User's Perspective.* Clarendon Press: Oxford.

KSHIRSAGAR, AM (1972) *Multivariate Analysis.* Marcel Dekker: New York.

LANKFORD, PM (1969) Regionalization: Theory and alternative algorithms. *Geogr. Anal.* **1**, 196-212.

LEE, RCT, SLAGLE, JR AND BLUM, H (1977) A triangulation method for the sequential mapping of points from N-space to two-space . *IEEE, Trans. Comput.* **C-26** 288-292.

LEHMANN, EL (1975) *Nonparametrics: Statistical Methods Based on Ranks.* McGraw Hill: San Francisco.

LEFKOVITCH, LP (1984) A nonparametric method for comparing dissimilarity matrices, a general measure of biogeographical distance, and their application. *Amer. Natur.* **123**, 484-499.

LEFKOVITCH, LP (1985) Further nonparametric tests for comparing dissimilarity matrices based on the relative neighborhood graph. *Math. Biosci.* **73**, 71-88.

LING, F (1972) A probability theory for cluster analysis. *J. Amer. Statist. Ass.* **68**, 159-164.

LUBISCHEW, AA (1962) On the use of discriminant function in taxonomy. *Biometrics* **18** 455-477.

MANTEL, N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209-220.

MANTEL, N AND VALAND, JC (1970) A class of permuational and multinomial tests arising in epidemiological research *Biometrics* **26**, 687-700.

MARDIA, KV (1970) Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519-530

MARDIA, KV AND ZEMROCH, PJ (1975) Algorithm AS84. Multivariate measures of skewness and kurtosis. *Appl. Stat.* **24**, 262-265.

MARDIA, KV, KENT, JT AND BIBBY, JM (1979) *Multivariate Analysis.* Academic Press: London.

MATULA, DW AND SOKAL, RR (1980) Properties of the Gabriel graphs relevant to geographic variation research and the clustering of points in the plane. *Geogr. Anal.*, **12**, 205-222.

MEAD, R (1966) A relationship between individual plant spacing and yield. *Ann. Bot. (New Series)* **30**, 301-309.

MIELKE, PW, BERRY, KJ AND JOHNSON, ES (1976) Multi-Response Permutation Procedures for *a priori* classifications. *Comm. Stat. Theory and Methods* , **5**, 1409-1424.

MIELKE, PW (1978) Clarification and appropiate inferences for Mantel and Valand's nonparametric multivariate analysis technique. *Biometrics*, **34**, 277-282.

MIELKE, PW (1979) On asymptotic non-normality of null distributions of MRPP statistics. *Comm. Stat. Theory and Methods* , **8**, 1541-1550

MIELKE, PW, BERRY, KJ., BROCKWELL, PJ AND WILLIAMS, JS (1981) A class of nonparametric tests based on multi-response permutation procedures. *Biometrika* **68**, 720-724

MIJARES, TA (1990) The normal approximation to the Bartlett-Nanda-Pillai trace test in multivariate analysis. *Biometrika* **77**, 230-233.

MOOD, AM (1940) The distribution theory of runs. *Ann. Math. Stat.* **11**, 367-392

MOON, JW (1970) *Counting Labelled Trees.* Canadian Mathematical Monographs, No 1. Canadian Mathematical Congress, Montreal.

MORAN, PAP (1948) The interpretation of statistical maps. *J. Roy. Statist. Soc. B*, **10**, 243-251.

MOSES LE AND OAKFORD, RV (1963) *Tables of Random Permutations.* Stanford University Press: Stanford.

O'ROURKE, J (1982) Computing the relative neighbourhood graph in the L_1 and L_∞ metrics. *Patt. Recogn.*, **15**, 189-192.

PICKARD, DK (1982) Isolated nearest neighbours. *J. Appl. Prob.* **19**, 444-449.

PREPARATA, FP AND SHAMOS, IM (1985) *Computational Geometry: an Introduction.* Springer Verlag: New York.

PRIM, RC (1957) Shortest connection networks and some generalizations. *Bell System Tech. J.* **36**, 1389-1401.

PURI, ML (1964) Asymptotic efficiency of a class of c-sample tests. *Ann. Math. Stat.* **35**, 102-121.

PURI, ML AND SEN, PK (1971) *Nonparametric Methods in Multivariate Analysis.* John Wiley and sons: New York.

QUINE, MP AND WATSON, DF (1984) Radial generation of n-dimensional Poisson processes. *J. Appl. Prob.* **21**, 548-557.

REEVE, ECR (1941) A statistical analysis of taxonomic differences within the Genus *Tamandua* Gray (Xenartrha). *Proc. London Zool. Soc., A.* **111**, 279-302.

RIPLEY, BD (1979) Tests of 'randomness' for spatial point patterns. *J. Roy. Statist. Soc. B* **41**, 368-374.

RIPLEY, BD (1981) *Spatial Statistics.* John Wiley and sons: New York.

RIPLEY, BD (1987) *Stochastic Simulation.* John Wiley and sons: New York.

ROBERTS, FDK (1968) Random minimal trees. *Biometrika* **55**,255-258

ROBERTS, FDK (1969) Nearest neighbours in a Poisson ensemble. *Biometrika* **56**, 401-406

ROBINSON, T (1987) *A k-sample Version of the Friedman-Rafsky Multivariate Runs Test.* Unpublished Manuscript. University of Western Australia.

ROHLF, FJ (1975) Generalization of the gap test for the detection of multivariate outliers. *Biometrics* **31**, 93-101.

ROSS, GJS (1969) Algorithm AS 13: minimum spanning tree. *Appl. Stat.* **18**, 103-104.

SAMMON, JW (1969) A non linear mapping for data structure analysis. *IEEE Trans. Comput.* **C18**, 401-409.

SCHILLING, MF (1986) Multivariate two-sample tests based on nearest neighbours. *J. Amer. Statist. Ass.* **81** 799-806.

SCHOLZ, FW AND STEPHENS, MA (1987). *K*-sample Anderson-Darling tests. *J. Amer. Statist. Ass.* **82**, 918-924.

SCHWERTMAN, NC (1982) Algorithm AS174: multivariate multisample nonparametric tests. *Appl. Stat.* **31**, 80-85

SEAL, HL (1968) *Multivariate Analysis for Biologists.* Methuen: London.

SEBER, GAF (1984) *Multivariate Observations.* John Wiley and sons: New York.

SIEDLECKI, W, SIEDLECKA, K AND SKLANSKY, J (1988a) An overview of mapping techniques for exploratory pattern analysis. *Patt. Recogn.*, **21-5**, 411-429.

SIEDLECKI, W, SIEDLECKA, K AND SKLANSKY, J (1988b) Experiments on mapping techniques for exploratory pattern analysis. *Patt. Recogn.*, **21-5**, 431-438.

SIEMIATYCKI, J (1978) Mantel's space-time clustering statistic: computing higher moments and a comparison of various data transforms. *J. Statist. Comput. Simul.* **7**, 13-31.

SIBSON, R (1978) Locally equiangular triangulations. *Comput. J.* **21** 243-245.

SIBSON, R (1980) The Dirichlet tessellation as an aid in data analysis. Scand. J. Stat. **7**, 14-20.

SIBSON, R (1981) *TILE4 Manual.* University of Bath.

SIBSON, R (1984) Present position and potential development: some personal views. Multivariate Analysis. *J. Roy. Statist. Soc. A* **147**, 198-207.

SILVERMAN, BW (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall: London.

SIMAIKA, JB (1941) On an optimum property of two important statistical tests. *Biometrika* **32**, 70-80.

SMALL, E AND LEFKOVITCH, LP (1986) Relationships among morphology, geography and interfertility in *Medicago*. *Can. J. Bot.* **64** 45-52.

SMALL, GC (1990) A survey of multivariate medians. *Int. Statist. Rev.* **58**, 263-285.

SMIRNOV, NV (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Moscow Univ.* **2**, 3-6.

SMITH, K (1953) Distribution-free statistical methods and the concept of power efficiency. In *Research Methods in the Behavioral Sciences.* (L. Festinger and D. Katz, eds.), pp.536-577. Dryden, New York.

SMITH, SP AND JAIN, AK (1984) Testing for uniformity in multidimensional data. *IEEE Trans. Patt. Anal. Mach. Intel.* **PAMI-6**, 73-80.

STEELE, JM (1980) Shortest paths through pseudo-random points in the d-cube. *Proc. Amer. Math. Soc.* **80**, 130-134.

STEELE, JM (1988) Growth rates of Euclidean minimal spanning trees with power weighted edges. *Ann. Prob.* **16** 1767-1787.

STEELE, JM, SHEPP, LA AND EDDY, WF (1987) On the number of leaves of a Euclidean minimum spanning tree. *J. Appl. Prob.* **24**, 809-826.

SUPOWIT, KJ (1983) The relative neighbourhood graph with an application to minimum spanning trees. *J. ACM* **30**, 428-447.

TAYLOR, P (1987) Contribution to the discussion of the paper by Jones and Sibson. *J. Roy. Statist. Soc. A* **150**, 32-33.

TOUSSAINT, GT (1980) The relative neighbourhood graph of a finite planar set. *Patt. Recogn.* **12**, 261-268.

TRACY, DS AND TAJUDDIN, IH (1985) Extended moment results for improving inferences based on *MRPP*. *Comm. Stat. Theory and Methods* **14**, 1485-1496.

TRACY, DS AND TAJUDDIN, IH (1986) Empirical power comparisons of two MRPP rank tests. *Comm. Stat. Theory and Methods* **15**, 551-570.

TRACY, DS AND KHAN, A (1987) MRPP test in L_1 norm. *Comput. Statist. Data Anal.* **5**, 373-380.

UPTON GJG AND FINGLETON, B (1985) *Spatial Data Analysis by Example, vol I: point pattern and quantitative data.* John Wiley and sons: New York.

URQUHART, RB (1980) Algorithms for computation of relative neighbourhood graph. *Electr. Lett.* **16**, 556-557.

URQUHART, RB (1982) Graph theoretical clustering based on limited neighbourhood sets. *Patt. Recogn.* **15**, 173-187.

WALD, A AND WOLFOWITZ, J (1940) On a test whether two samples are from the same population. *Ann. Math. Stat.* **11**, 147-162.

WATSON DF (1981) Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *Comput. J.* **24**, 167-172.

WHALEY, FS (1983) The equivalence of three independently derived permutation procedures for testing the homogeneity of multidimensional samples. *Biometrics* **39**, 741-745.

WHALEY, FS AND QUADE, D (1985) Optimizing the power of the two sample multidimensional runs statistic: guidelines based on computer simulation. *Comm. Stat. Theory and Methods* **14**, 1-11.

WHITNEY, VKM (1972) Minimal spanning tree. *Comm. ACM* **15**, 273.

WICHMANN AB AND HILL, ID (1982) Algorithm AS183: an efficient and portable pseudo-random number generator. *Appl. Stat.* **31**, 188-191

WILSON, RJ (1972) *Introduction to Graph Theory.* Oliver and Boyd: Edinburgh.

YAO, AC (1982) On constructing minimum spanning trees in k-dimensional spaces and related problems. *SIAM J. Comput.* **11**, 721-736.

ZAHN, CT (1971) Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE, Trans. Comput.* **C20**, 68-86.

ZIMMERMAN, GM, GOETZ, H AND MIELKE, PW (1985) Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology* **66** , 606-611.

MJFCB fecit — Bath, 10/I/1992.